POLITECNICO DI TORINO

Master Degree course in Artifical Intelligence and Data Analytics

Master Degree Thesis

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

Supervisors
Prof. Santa Di Cataldo
Prof. Francesco Ponzio

Candidate
Stiven Hidri

Academic Year 2023-2024

Acknowledgements

I will never thank my parents enough, who, despite the difficulties throughout these years, have always supported me. They asked me for nothing, besides commitment.

I will never thank my little brother enough, who, by looking up to me, pushed me to be better. Yet, I feel like there is still a long way to go.

I will never thank my best friends enough, the ones that I've known forever and the ones that I don't, but feel like it. We have each other's backs.

My most sincere acknowledgments go to both my supervisors for helping me during this work. Thanks to you, I believe I found the field I'd like to work in one day.

Finally, thank you for showing so much strength in such a difficult moment. Working on this topic surely felt different in these last months.

List of Figures

List of Tables

Contents

Abstract

Gamma Knife is an advanced form of radiation therapy allowing non-invasive tumor treatment by delivering high radiations doses to localized brain regions. Early detection of tumor progression after Gamma Knife radiation therapy can significantly influence treatment decisions and improve outcomes in brain cancer patients. In this work we propose several deep learning solutions to predict local progression of metastatic brain tumors after Gamma Knife radiation therapy exploiting MRI scans, radiation dose maps and clinical patient information. The training set consists of 140 stable and 13 recurrent lesions, while the test set contains 81 stable and 10 recurrent lesions. Oversampling, focal loss, data augmentation and thresholding are adopted to face data imbalance along with cross validation to improve model generalization. We explored different multimodal fusion techniques: input fusion via Discrete Wavelet Transform, concatenation-based information fusion, LSTM-based information fusion and attention-based fusion. Despite the challenges posed by data scarcity and imbalance, our proposed solutions demonstrated the potential of multimodal deep learning models leveraging MRI, Dose and clinical information, laying the foundations for future studies aimed at timely, accurate tumor progression predictions to improve life quality and expectancy of patients affected by metastatic brain cancer.

Chapter 1

Introduction

1.1 Brain tumors

Tumors are abnormal growths of cells. They can labeled as benignant or malignant (cancer) wether they can spread or invade nearby tissues, or even metastasize in other parts of the body. Specifically, brain tumors can cause several health problems by applying pressure to critical structures, potentially disrupting essential functions, and often proven difficult to fully remove.

The latest survey conducted by the Italian Oncological Medical Association [6] in 2023 reveals that in Italy around 52.800 people have been diagnosed with central nervous system cancer: in 2022 around 6300 new cases have been discovered, 3600 being males and 2700 being females. Expectancy of life in the first 5 years is 24% in males and 27% in females. Moreover, meanwhile the trend in Italy in the last few years appears to be stable, many industrialized countries, such as USA and UK, are witnessing a constant increment of incidence rate.

Metastatic brain tumors are much more frequent than primary ones [7]: the study conducted by Saha et al. [8] in 2013 revealed a ratio of 10:1 with respect to primary brain tumors. Moreover, 25% of cancer patients develop metastasis in the brain.

1.2 Gamma knife

The first and most common treatment for brain tumors is surgery, which aims to remove as much of the tumor as possible while preserving surrounding healthy tissues. This approach helps relieve pressure on critical structures and may help other ongoing treatments. Gamma Knife (GK), on the other hand, is a non-invasive stereotactic radiosurgery (SRS) technique. It works by focusing multiple low-dose radiation beams precisely on the tumor's location. While within a single beam the radiation dose is minimal, the intersection of all beams at the tumor contains a much higher dose, effectively treating it without performing any incision. The outcome of this procedure cannot be detected immediately after, and follow-ups will determine the tumor's response to the treatment. The RANO-BM [9] criteria provides standardized guidelines for evaluating the response

of brain metastases after the therapy. The 4 response categories described in the guidelines are: Complete Response (CR), Partial Response (PR), Stable Disease (SD) and Progressive Disease (PD).

Accurate estimate of SRS outcome would enable effective and decisive adjustments to the treatment planning and design of metastatic brain cancer patients, ultimately improving health outcomes and quality of life. Moreover, narrowing down the treatment outcome estimation locally to individual lesions could further optimize personalized treatments and substantially improve prognostic accuracy.

1.3 Medical domain information

1.3.1 Imaging

In the medical domain different types of scans are used to reveal important characteristics which allow to correctly diagnose, classify, and monitor diseases. The most popular ones are:

- **Magnetic resonance imaging (MRI)**: is one of the most commonly used imaging for brain lesion analysis. A key factor is its ability to allow visualization on all planes (axial, sagittal and coronal). The most common types are:
  - **T1-weighted**: it can also be performed while infusing gadolinium. By resulting brighter in the scan it allows to enhance lesions with respect to surrounding healthy issues.
  - **T2-weighted**
  - **Fluid Attenuated Inversion Recovery (FLAIR)**
  - **Diffusion-Weighted Imaging (DWI)**
- **Computed Tomography (CT)**
- **Positron Emission Tomography (PET)**
- **Planned radiation dose map**: in this work we will exploit the planned radiation therapy produced by the software that plans the GK treatment. It allows to have a 3D representation on where the radiation dose will be deposited and how much.

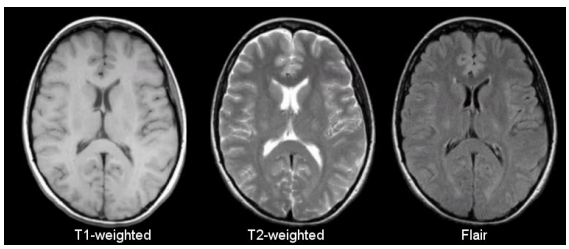

Figure 1.1: From left to right, T1W, T2W and FLAIR MRI scans. [1]

1.3.2 Clinical data

Clinical data complements imaging, allowing to introduce context for decision making. Demographics is usually included, which can be more or less determinant depending on the specific task. Primary tumor location and tumor histology allow to describe, in the case of metastatic tumors, the original cancer location and its cell structure. Lesion location allow to identify in which region the tumor has developed. Number of metastases and longest diameter are relevant information, which can help the prognosis of treatment outcome.

More specific clinical data are:

- Graded Prognostic Assessment (GPA): is a scoring system used to stratify patients into prognostics groups. The higher the GPA score the better the prognosis. Some of the considered factors are primary tumor type, age, KPS score and number of brain metastases.

- Karnofsky Performance Status (KPS): is a score assessing the patient's functional status. The higher the score the more the patient functional abilities are uncompromised.

- Edema Index (EI): is a quantitative measure of the swelling developed around the lesion due to fluid accumulation.

1.4 Machine learning within the medical domain

Machine Learning based solutions have demonstrated great capabilities in extracting relevant and discriminative correlations between medical imaging and clinical data, enabling applications such as disease classification, prognosis prediction and treatment planning.

Traditional models, like SVM, Decision Trees and Logistic Regression have been largely used in the medical domain, even though they require manual feature engineering. Deep learning, instead, allows to automatize this stage given its ability to directly process raw and complex information (images, videos, text, audio) delivering often higher performances.

Convolutional Neural Networks (CNN) are a popular choice for image classification tasks since they excel at learning hierarchical features; starting from simpler edges and texture, in shallower layers, to more complex patterns, like shapes and objects, in deeper layers. Moreover, many pretrained models are available, which allow to extract relevant features with limited samples through transfer learning. Although Recurrent Neural Networks are designed to handle sequential data, they have been proven capable of extracting important spatial correlations within 3D medical imaging. For what concern transformers, they have recently found their application in medical imaging thanks to the development of Visual Transformers (ViT): by exploiting the self-attention mechanism they manage to capture global relationships, a fundamental characteristic which can be determinant in multimodal information contexts.

1.5 Multimodal information

It is not uncommon, within the medical domain, the availability of multiple input modalities: it is crucial to find effective ways to leverage them all to maximize models performances by exploiting complementary information. Information fusion can be adopted at different levels within the learning pipeline. Input fusion, fuses or concatenates the different inputs before the deep learning backbone is reached. Intermediate fusion happens whenever the features, extracted by some deep learning backbones, are merged or concatenated before the final classifier. Output fusion, on the other hand, tries to combine the classification outputs of different classifiers, which have independently classified each input modality, to produce a final prediction. In recent years, attention-based fusion techniques have gained a lot of popularity [2]: they exploit attention relationships among modalities to perform feature fusion.

1.6 Data scarcity and imbalance

Nevertheless, a lot of data is needed to produce good results. Yet, medical imaging gathering is a slow and costly job: it requires time, approval requests, specialized personnel and resources. These constraints often lead to datasets containing limited amount of samples. Moreover, some domains are intrinsically imbalanced which negatively affect deep learning models performances. Ad hoc strategies have been developed to directly face data imbalance and scarcity. To increase the number of samples, the first step consists of augmenting the dataset by applying augmentation techniques like rotation, flipping, shearing, and many others. More advanced technique, like Generative Adversarial Networks (GANs), allow to generate augmented samples, by learning intrinsic structural information from the available ones. For what concerns data imbalance, it can be dealt by choosing specific loss functions. Binary weighted cross entropy allows the model to focus more on minority classes by assigning different weights to the classified samples, based on the class they belong to. Focal loss [3], instead, allows the model to focus on harder to classify samples (in other words, misclassified samples). Cross validation is very effective when it comes to hyper-parameters tuning, but it also help to maximize the use of limited samples: by dividing the dataset into k folds ensures that every part of the dataset gets used in both training and validation or test phases, across multiple iterations. Models trained on different training sets are produced, which could also be used together with straightforward techniques, such as majority voting or prediction averaging, in order to maximize performances.

1.7 Proposed work

In this thesis, we will address the task of predicting local tumor recurrence after Gamma Knife treatment by exploiting pre-treatment MRI scans, ROI masks, planned radiation dose maps, and clinical information of patients affected by metastatic brain tumors. Tumors that shrink or stabilize after the treatment are labeled as 'stable', while remaining

cases are labeled as 'recurrent'. To address the challenges posed by the dataset's imbalance, focal loss, data augmentation, and thresholding techniques are employed to mitigate its effects. To leverage all input modalities, different multimodal fusion techniques are tested. The model proposed alongside the dataset, extracts the features from the imaging through convolutional layers and finally concatenates them with the clinical features. Then, an input fusion technique is developed, which makes use of Discrete Wavelet transform to fuse MRI scans and radiation dose maps. The resulting input is then fed to a convolutional backbone to extract relevant features. Additionally, a state of the art model is adapted into our domain: it makes use of a CNN backbone to extract local features from overlapping 2D slices of both MRI scans and planned radiation dose maps. Finally, an LSTM network is used to include long-range relationships between different slices. Lastly, an attention-based fusion architecture is again adapted to our domain: it exploits CNN for local feature extraction from 2D slices and self-attention mechanism to find important correlations between data modalities. Despite not reaching satisfactory results, important considerations are made for future works, regarding both new techniques and future datasets creation.

Chapter 2

Related works

2.1 Current state of the art

2.1.1 MRI radiomics features and clinical data

Du et al. [10] developeded several machine learning solutions that make use of MRI radiomics and clinical risk factors. Patients who participated in the study, 337 in total, had no more than four brain metastases, and had complete clinical data and pretreatment imaging.

Pre-treatment MRI scans included Axial CE T1WI, T2WI and DWI acquired on 1.5T MRI system. In addition to MRI scans, whole tumor (enhancement and non-enhancement) and peritumoral edema ROIs masks were also produced. Clinical features included in the study were: Graded Prognostic Assessment (GPA), number of metastases, KPS edema index, primary tumor type, lesion location, age, tumor volume and sex. The response criteria follows the RANO-BM criteria: specifically in this work complete response (CR) and partial response (PR) are labeled as locally effective (LE), stable disease (SD) is labeled as Locally Stable (LS) and progressive disease (PD) as locally ineffective (LIE).

Radiomics features were extracted by applying different image filters to all combinations of imaging sequences and regions of interest. An exhaustive feature selection has been conducted to choose only the most relevant and unique features. The majority of chosen features, 233 in total, were texture and wavelet, demonstrating the relevance of CE T1W sequence, tumor ROIs, advanced radiomics textures, and wavelet image features.

They analyzed clinical features by computing Spearman correlation coefficient matrix and applying logistic regression. They ended up finding number of brain metastasis and GPA the most discriminant features. Follow along KPS score and edema index, while sex and primary tumor type did not appear to be relevant for the task.

The six chosen classifiers were Gaussian Naive Bayesian (GNB), k-Nearest Neighbors (KNN), random forest (RF), adaptive boosting (AB), support vector machine (SVM) with linear kernel, and multilayer perceptron (MLP).
Experiments were run adopting both and alternatively radiomics and clinical data. Classifiers fed with both modalities delivered the best performances, with SVM being the most accurate classifier reaching above 0.9% AUC in all experiments.

2.1.2 CT scans and CNN

The work proposed by Cha et al. [11] addresses SRS outcome prediction by adopting computer tomography (CT) imaging and CNN. In total 89 patients and 110 tumors were considered for this work. Complete (CR) and partial (PR) response were labeled as 'responder' whereas stable (SD) and progressive (PD) were considered as 'non-responder'. An additional note is that tumors with a longest diameter larger than 35mm or smaller than 10mm were not included in this work. Gathered CT imaging consisted of one axial slice for each tumor, including its central point with marginal surrounding tissue. The CNN structure involved 2 convolutional and pooling layers and two final fully connected layers. 50 different datasets were created, divided into training, validation, and evaluation groups. Five groups of 10 different datasets each were instantiated, ensuring the same evaluation datasets across groups. An additional ensemble model averages the prediction of all models within each group. The AUC for individual CNN models ranged from 0.602 to 0.8262. The ensemble models showed improved performance with AUCs ranging from 0.761 to 0.856.

2.1.3 Recurrent Neural Networks

The study conducted by Cho et al. [12] proposes 3 solutions, namely maximum axial diameter (Dmax), radiomics, CNN, and CNN-based GRU to predict SRS outcome by exploiting MRI scans.

The dataset included 194 patients affected by brain metastasis, collecting 369 individual lesions. Train and test split were created by randomly splitting the whole dataset with a ratio of 8:2. An additional external validation set was employed, containing 43 patients with 62 total target lesions. The treatment outcomes were categorized between progressive disease 'PD' and 'non-PD', which includes Complete response (CR), partial response (PR), and stable disease (SD). MRI imaging included always a pre-treatment scan, and up to 3 follow-up scans. Again, a lower bound of at least one dimension being higher then 5mm has been set to avoid analyzing too small lesions.

The maximum diameter of each brain metastasis was measured on the representative axial plane, and radiomics features were extracted through the PyRadiomics library. Both information were analyzed by XGBoost models. The simple CNN architecture extracts the features from the 4 scans inputs and concatenates them. The CNN-based GRU model splits into a first version, where features from 3D MRIs are extracted through a 3D ResNet34 backbone, and a second one, where 3 patches, one for each orthogonal plane, are extracted and fed to a 2D ResNet34 backbone. In both versions, extracted features are then fed to a GRU network. Optimal thresholds were calculated by considering Youden's J statistic and ROC. The 2D Conv-GRU model outperformed al the others by reaching a mean AUC 0.8782 in the test set and 0.8341 in the external validation set. Recorded results showed also that accuracy increases with the number of follow-up images used.

RNNs were explored also by Jalalifar et al. [13] for the purpose of local failure prediction after SRT. The dataset included 99 patients presenting a total of 116 lesions. Imaging consisted of T1w and T2-FLAIR scans, collected prior to the SRT treatment. An independent set of 25 patients and 40 target lesions was used for final evaluations.

Complete Response (CP), Partial Response (PR), or Stable Disease (SD) outcomes were labeled as Local Control (LC) whereas Progressive Disease (PD) was labeled as Local Failure (LF).

The proposed solution consists of a InceptionV2ResNet network used as backbone for 2D slices feature extraction of T1w and T2-FLAIR scans. Features extracted were processed by LSTM, Seq2Seq and transformer networks to incorporate spatial dependencies between slices. Additionally, clinical features were also taken into consideration: an internal study defined as most discriminative for the task histology, tumor location, tumor size, and number of brain metastases. The architecture exploiting the LSTM network will be explained thoroughly in the methods sections since it has been replicated to exploit radiation dose maps and MRI imaging in our work. The configuration using the LSTM network with clinical features reached an AUC on the independent test set of 0.86, the highest among all solutions.

2.2 Multimodal medical fusion

In the medical domain more than often is possible to collect data from different sources carrying complementary information: medical imaging, clinical records, and lab results just to name a few. Moreover, significantly in neurology, medical imaging is able to give non-invasive and detailed insights for various purposes like diagnosis of neurological disorders and disease monitoring. It is then essential to experiment and develop new strategies that are able to find and exploit correlation between different modalities. We will focus prevalently into multimodal image fusion given the nature of the information that we will treat. To further support the importance of multimodal image fusion in the medical domain it is enough to look at the trend of the number of published articles in recent years dealing about it 2.1. Different fusion strategies can be classified based on

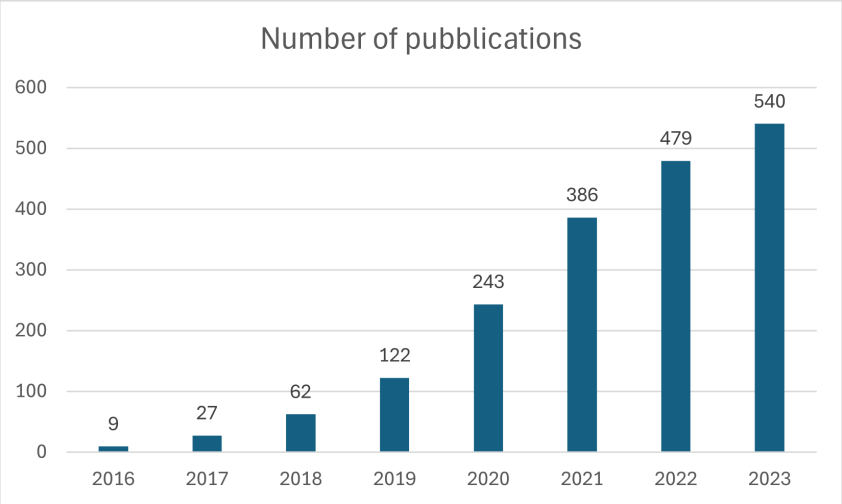

Figure 2.1: Number of pubblications regarding multimodal fusion of medical imaging on PubMed from 2016 to 2023. [2]

when the information fusion is performed during the classification pipeline and which strategy is used to combine different information.

2.2.1 Input fusion

Input fusion allows to combine data from multiple imaging sources before accessing the deep learning backbone. The simplest way consists of concatenating the input images along one dimension. Otherwise, it is possible use advanced methods to merge the different input modalities pixel by pixel, or eventually, voxel by voxel in the case of 3D images.

The interesting work conducted by Rallabandi and Seetharaman [14] proposes an Inception-ResNet wrapper model for automated detection of dementia stages using combined MRI and PET scans. Here, fusion between MRI and PET imaging was achieved using 2D Fourier and Discrete Wavelet Transform (DWT). This work gave us the inspiration for developing a much simpler DWT-based fusion technique which will be further discussed in the Methods section.

Input fusion usually requires the inputs to have the same structure and spatiality. In many cases, different image modalities are retrieved together at the time of clinical photography and may already share the same coordinates and spatial characteristics. If this is not the case, resampling must be applied which allows to accurately overlay images taken in different conditions.

2.2.2 Intermediate fusion

Intermediate fusion consists of combining data of different modalities at an intermediate level, typically after applying preprocessing. Here the features from different modalities are extracted through deep learning backbones and then are merged or concatenated. An improvement of this technique is called hierarchical fusion, which allow to fuse the extracted features at different layers allowing to progressively combine information from different sources, starting with simple combinations and gradually moving to more complex interactions. A different application of intermediate fusion consists of using an additional deep learning backbone in order to fuse the previously extracted features.

The base model [4], provided within the article dealing about the dataset we used, applies an intermediate fusion technique, involving the concatenation of features extracted from MRI scans, radiation dose maps and clinical data. On the other hand, the models adopting recurrent neural network by Jalalifar et al. [13] and Cho et al. [12] exploit LSTM and GRU networks to fuse the features coming from different images.

2.2.3 Attention-based fusion

The advent of transformers has revolutionized the field of Neural Language Processing (NLP) by moving away from traditional RNNs and CNNs to analyze sequence of information. One of the main advantages is surely the more scalable and parallelizable architecture. In recent years, transformers have been explored in a variety of other domains, like image classification. Multimodal transformer solution have been developed to exploit the Self-Attention mechanism to find relevant correlations between different

data modalities. Among many strategies, the work proposed by Dai, Gao and Liu [5] describes TransMed, an attention-based fusion solution which efficiently captures both local features (via CNNs) and long-range dependencies (via transformers) between different imaging modalities which makes it a valuable solution for limited dataset. This strategy has been adapted to our treatment outcome prediction task and will be further discussed in the Methods section.

Chapter 3

Materials and methods

3.1 Dataset

Wang et al. [4] released a brain cancer MRI dataset along with Gamma Knife treatment planning and follow-up information in order to promote the development of machine learning solutions able to predict local tumor recurrence given the current treatment planning, MRI imaging, and patient clinical information. In total, information was retrieved from 47 patients affected by metastatic brain tumors resulting in 244 lesions with annotations. Patients might have undergone from 1 up to 8 courses. Multiple courses might have been needed in case of local recurrence occurred or new metastatic tumors developed in the brain. Local lesion progression was identified by follow-ups: "stable" includes Complete Response (CR), Partial Response (PR) and No Change (NC). "recurrence", instead, indicates local progression: eventual Beside or Remote Recurrence are treated as new lesions.

3.1.1 Structure

Each subject is identified by a unique id and each patient's course is identified by a progressive number. For each course are available: the MRI series recorded in preparation for the treatment, the RTStruct information containing the ROI masks and the planned radiation dose maps. Fig. 3.1 shows the file structure map. Moreover, primary tumor location, tumor histology, age at diagnosis and gender are recorded for each course. Lesion level information is stored too: it contains for each lesion the location, the label ("stable" or "recurrence"), fractions (number of session in which the total radiation dose was delivered for that course) and finally the number of months elapsed between the treatment and follow-up imaging. Fig. 3.2 shows some examples of clinical records and how they are managed across different tables.

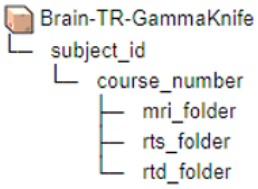

Figure 3.1: Dataset file structure.

Unique patient ID	Course number	Primary tumor location	Histology	Age at diagnosis	Gender
463	1	Brain Mets-Lung	Adenocarcinoma of the lung	60	Male
463	2	Brain Mets-Lung	Adenocarcinoma of the lung	60	Male
247	1	Brain Mets Breast	Invasive ductal carcinoma of Rt Breast	63	Female
408	1	Brain Mets Lung	Adenocarcinoma of the lung	64	Male
...	...	...	...	...	...

Unique patient ID	Course number	Lesion number	Lesion location	MRI type	Duration (months)	Fractions
463	1	1	Lt Frontal	recurrence	11	1
463	2	2	R Motor Cortex	stable	8	1
463	2	3	Lt Post Temporal	stable	8	1
463	2	4	Lt Lat Cerebellum	stable	8	1
...	...	...	...	...	...	...

Figure 3.2: Course and lesion level records.

3.1.2 MRI scans, ROI masks and planned radiation dose maps

MRI scans are T1 MPRAGE with gadolinium contrast. The original axial plane is 256x256 pixels while the slice thickness varies individually.

Lesion mask delineation was a collaborative effort involving at least 1 radiation oncologist, 1 neurosurgeon, and 1 neuroradiologist. The delineation limit criterion was generally determined based on absorbed contrast, including regions identified as suspicious by consensus.

Radiation dose maps are computed by the treatment planning system (TMR10 algorithm of the GammaPlan software) which does not include the heterogeneity of the tissues within calculations. Moreover, no uncertainties regarding the accuracy of the delivery are included.

Fig. 3.3 shows some slice samples from the whole imaging of subject 427 regarding its second course.

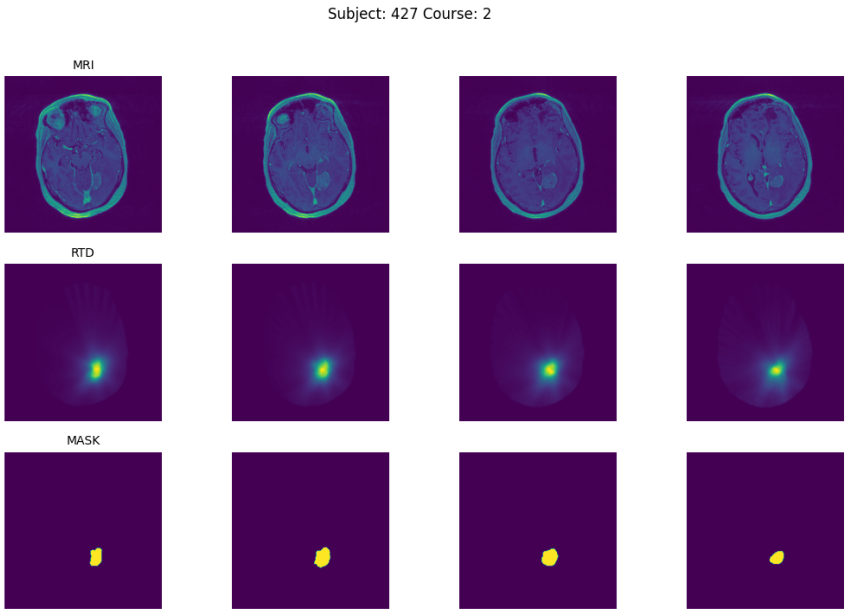

Figure 3.3: Slice samples from the whole imaging. From the top to the bottomo row we have respectively MRI, planned radiation dose map and ROI mask.

3.1.3 Reading pipeline and preprocessing

The raw DICOM files containing all the imaging weigh overall around 10GB. The remote server used for running all the tests had limited disk space and it was not possible to store directly all the data there. Moreover different trials have been carried out before choosing the final preprocessing strategies and it was not possible to read, process, and save each time all the information. For this reasons, the reading pipeline was divided

into two stages: the first stage consists into extracting information regarding individual lesions from MRI and radiation dose maps by exploiting ROI masks.

As also stated in the article, MRI scans and radiation dose maps share both incompatible spacing and different dimensions. Moreover, different MRI scans presented different spacing as well, thus a general resampling was needed in order to proceed. These steps are not trivial, but fortunately the powerful library SimpleITK [15] is able to simplify them.

Medical images processed by this library store metadata along with the image to define how the voxel information are mapped into the physical world. The region in the physical space that an image occupies is defined by the origin, the spacing, the size and the direction cosine matrix. The origin defines the coordinates of the voxel with index (0, 0, 0) in the world coordinate system (for example in millimeters). Spacing defines the distance between adjacent voxels along each dimension in real world unit of measure. For example a spacing of (1, 1, 0.5) means that voxels are 1mm apart along x and y axis and 0.5 mm along the z axis. The size defines the the discrete number of voxels along each dimension of the image. Finally, the direction cosine matrix contains the orientation of the image axes relative to the world coordinate system. An additional parameter, which does not directly influence the resampling process, is the slice thickness. It measures the actual thickness of the tissue imaged in one slice and it impacts the resolution and quality of the image.

The first step we took consisted of resampling the MRI scans into a isotropic spacing of (1, 1, 1). Then, since the ROI masks share the original spacing of the MRI scan, they are resampled too, by taking as reference the resampled MRI scans. SimpleITK is able to align the two images by calculating the transformation matrix from the source and destination parameters described above. Then, interpolation is able to estimate the voxel intensity of transformed coordinates who fall between voxels. We adopted linear interpolation, which it is a popular choice given its ability to deliver a good compromise between quality and computational time. We used instead nearest neighbor interpolation when resampling ROI masks, since these are binary images where each voxel is either part of the ROI (value of 1) or not (value of 0). This avoids intermediate values, maintaining the integrity of the binary mask. Finally, each radiation dose map is resampled by using again as reference the resampled MRI scan. This process ensures that all the medical images share the same spacing of (1, 1, 1) and that each triplet of ROI masks, MRI scan and radiation dose map are aligned.

Each MRI scan contains different ROI masks, depending on the number of present lesions. These, unfortunately, are identified with a mixture between lesion location and progressive lesion numbering with inhomogeneous termed used across different masks. In order to associate each ROI to the clinical information, which contains the label of the lesion, string matching between the ROI identifier and the corresponding clinical record is performed by measuring the Levenshtein distance between the lesion location field, contained in the clinical data, and the ROI identifier. This process has been reviewed manually sample by sample. This was the only way to proceed given the lack of a direct and unique joining field between clinical records and the respective ROI mask identifier.

Then, after the association, all the ROI masks have been applied to the relative MRI

scans and radiation dose maps: a 5mm additional area around the mask was kept, since it has been shown to positively contribute to treatment outcome prediction [13] [16].

A final python dictionary stores, for each target lesion, the preprocessed imaging and relative clinical data. Clinical data contains the subject id too, which will be used during the splitting phase of the dataset. The followed pipeline allows to isolate all the information for each target lesion, and is thought for developing machine learning algorithms able to detect treatment outcome by analyzing each target tumor independently.

Fig. 3.4 shows examples of preprocessed lesions with relative planned radiation dose map.

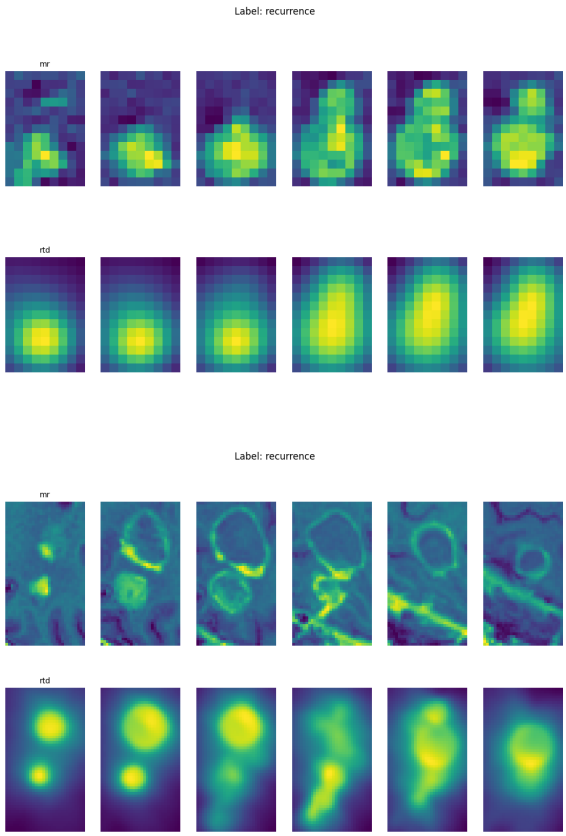

Figure 3.4: Stable lesion MRI and planned dose map.

The second stage is carried out in a second moment, during the initialization of the learning process. When the dataloader fires up, it reads the file structure containing all the extracted information and applies all the scaling operations. For what concerns the target lesions and the corresponding radiation dose maps, min-max scaling is applied which maps the voxel intensity values between 0 and 1. Minimum and maximum statistics are computed over the training set, in order to use them to process the validation and test samples. Moreover padding is added in order to even the input size of the preprocessed lesion MRI and radiation dose samples.

By running some statistics, it comes out that the minimum registered shapes are 11x12x13 and the maximum are 48x66x51. This means that smaller lesions will experience major padding, which can lead to worsening the performances. This will be further discussed in the Results section.

For what concerns the clinical data, thorough preprocessing steps have been followed too since, unfortunately, the available categorical values lack consistency in terminology and specificity. Moreover, as pointed out by other works on the same task, the available records may not be the most discriminant to detect the outcome of SRS treatment. Table 3.1 displays some examples of lesion locations. As it can be seen from the displayed values, some ROIs are more detailed than others and each term needs to be standardized.

L post inf temp	Crebellar culmen	L mesial cereb	R cereb hemisp
4 Lt Inf Cerebellar1	Rt Cerebellar	9 Lt Post Frontal 1	8 Rt Parietal 1
Lt ant inf frontal	Lt ParaMedian	Lt Ant Temp	Lt Frontal POle

Table 3.1: Examples of lesion locations.

An initial mapping between abbreviations and standard terms has been developed to standardize the records. After this primary step, the number of features obtained as a result of applying one hot encoding was still too high with respect to the number of available samples. Then the standardized lesion locations have been mapped to a more general value: specifically, each location has been mapped to its respective lobe and side (if present): parietal lobe, temporal lobe, occipital lobe, frontal lobe, cerebellum, and left or right side. If a lesion was not located in any particular lobe, the original location was left as it was . Table 3.2 shows some of these mappings.

Original lesion location	Generalized
Right postcentral	right parietal
Rt Sup Frontal	right frontal
Rt Medial Cerebellum	right cerebellar
Lt Frontal	left frontal
Rt Frontal deep	right frontal
Rt sup vermis	vermis
Rt Inf Med Cerebellu	right cerebellar
Rt Fourth Ventr	right ventricle
Lt Cerebellar	left cerebellar
CavityltCerebellar	left cerebellar

Table 3.2: Examples of mapping of ROIs to more general regions.

Primary tumor location and histology undergone to the same standardization process. The resulting input vector consists of 47 features. Input vector contains also information regarding the number of lesions in the current course and the lesion volumes. These information were not directly included but it was possible to retrieve them from the available clinical records.

3.2 Data imbalance

Table 3.3 shows the split used in the article proposing the dataet. Only 13 samples of the minority class were available during training and 10 for testing. The ratio of samples from the recurrence class with respect to the ones of the stable class is around 1:10. Ad hoc techniques must be taken into consideration in order to attenuate the effects of such imbalance.

Lesions	Stable	Recurrence	Total
Training set	140	13	153
Test set	81	19	91
Total	221	23	244

Table 3.3: Samples split overview.

The article proposing the dataset tries to balance the two classes before the training phase by oversampling the minority class through data augmentation. By rotating each lesion and the corresponding radiation dose maps by 3 angles (respectively 90, 180, 270) over each axis (x, y, z) they were able obtain such balancing. Other techniques are available in order to deal with imbalance, which can be obtained by also downsampling the majority class. Nevertheless, the loss functions can play a crucial role too: weighted binary cross-entropy introduces weights to the positive and negative classes, which induces the model to pay more attention to the minority class. Usually, the weight can be set as the inverse frequency of the class. Another loss function tailored for classification tasks in imbalanced datasets is focal loss [3]. Originally designed for object classification, it allows the model to focus on hard-to-classify examples by introducing a modulating factor to the cross-entropy loss.

$$-\alpha_t \cdot (1 - p_t)^\gamma \cdot \log(p_t) \tag{3.1}$$

Where:

$$p_t = \begin{cases} p & \text{if y} = 1 \\ 1 - p & \text{otherwise} \end{cases} \tag{3.2}$$

α is a weight term that enables to control the balance between the minority and majority class, ensuring that the minority class receives more focus. The modulating term $(1 - p_t)^\gamma$, instead, reduces the impact of well-classified samples (easy to classify): by increasing γ, we induce the model to concentrate more on misclassified samples (hard to classify). The graph depicted at Fig. 3.5, which was taken directly from the original article, shows the trend: the loss for well classified samples decreases significantly as the modulating term reaches 0.

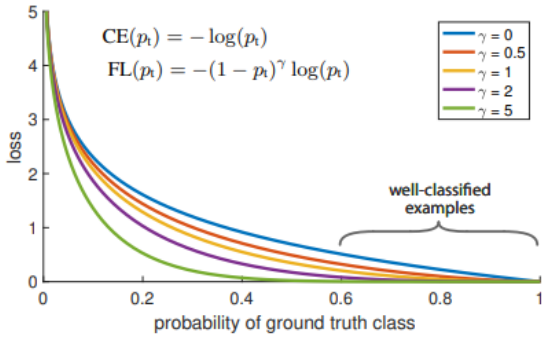

Figure 3.5: Focal loss trend. [3]

3.3 Data augmentation

Data augmentation allows to increase the diversity of the training set by applying different transformations to the original samples. It leads to improved generalization, thus reducing overfitting which is crucial when working with limited data. Moreover, it increases robustness, by allowing to handle variations like rotation or noise in real world data.

For this work, we decided to apply random rotation, random flip and random affine transformation. For what concerns random rotation, an angle between 0, 90, 180 and 270 was chosen along with a random axis (x, y or z). Random flip is performed randomly on one of the 3 axis with 0.5 probability. By applying random affine transformation we are able to apply a slight random rotation between -30 and 30 degrees, translation between -5 and 5 pixels and random scaling with a factor between 0.85 and 1.15.

3.4 Split criteria

Dataset splits are always performed by subject, and not by individual samples. If samples from the same subject were present in both train and test or validation set the model might exploit patterns unique to that subject rather then generalizing across individuals, which can be considered as data leakage. This practice is often used in the medical domain since it allows the model to evaluate on completely unseen subjects, mimicking real world scenarios. In this work we exploited stratified group splitting: it ensure the the distribution of the classes is preserved (stratification) while also ensuring the the same subject is not present of more than one set (grouping).

3.5 Cross validation

Cross validation is a useful tool that allows to exploit the limited amount of samples and tune hyper-parameters. Given the scarcity of samples of the recurrence class, only 23 in the whole dataset, we decided to perform 2 main cross validations in order to test

the performances of the proposed models: we apply a 10-fold cross validation on the entire dataset creating iteratively 10 folds, 9 for training and 1 for testing. From each training set a small portion of samples is held out for validation. This strategy allows to maximize the number of samples the model can process. Finally prediction results of each test set are concatenated and a global performance is computed. Secondly we perform a 5-fold cross validation only on the training set proposed by the dataset's article: at each iteration 4 folds are destined to the train set and 1 for the validation set.

3.6 Models

3.6.1 MLPCD

Clinical records, after the preprocessing steps, are embedded in an input vector containing one hot-encoded categorical values and min-max scaled discrete values. We implemented a Multi-Layer Perceptron (MLP) to produce discriminant embedding from the raw input. The MLP consists of 2 hidden layers, respectively of size 128 and 64. ReLU is chosen as activation function and both dropout and batch normalization are applied at each hidden layer. An additional final fully connected layer is added when classifying samples only using clinical records, in order to test the effectiveness of this strategy.

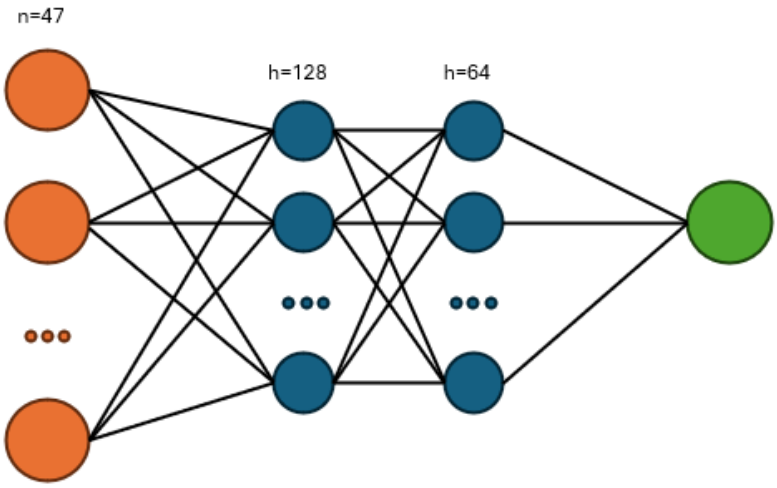

Figure 3.6: Multi layer perceptron for clinical records embedding.

The first hidden layer of higher dimensionality should allow to capture more complex non-linear relationships between clinical features. The second one, instead, refines the extracted features into a more compact representation while trying to preserve discriminative relationships.

3.6.2 Base model

Alongside the dataset, a novel model is proposed in order to produce a baseline evaluation which exploits intermediate fusion trough concatenation. The schema in figure 3.7 summarizes the model structure.

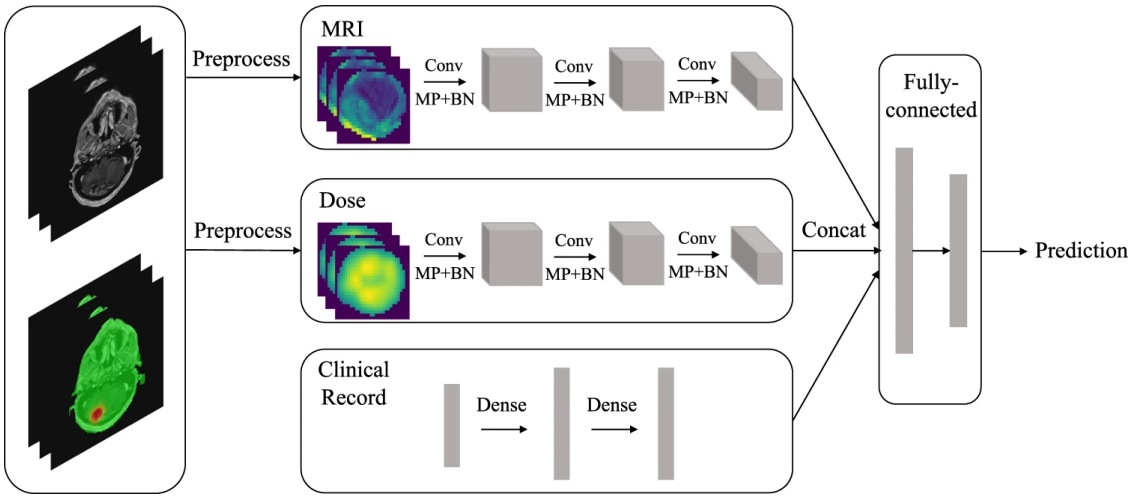

Figure 3.7: Overview of the baseline network included in the brain mri dataset. [4]

The model consists of 3 main branches, each one handling a separate data modality. High level features of both MRI and radiation dose maps are extracted through 3 sequential convolutional blocks consisting of one 3D convolutional layer, one max pooling layer, one batch normalization layer, and ReLU activation applied to the convolutional outputs. The final number of produced feature channels is 128. Finally, a global average pooling layer compacts the output to a flat feature vector of size 128. Then, two fully connected layers project the resulting feature vector to a feature size of 512 and back to a feature size of 128. ReLU activation function is employed along with dropout with the two fully connected layers.

For what concerns the clinical data, our proposed preprocessing steps results in a different input shape and so the branch handling clinical data has been replaced with the MLP CD model, stripped of its final fully connected layer. The resulting feature vectors from each branch are concatenated producing a final feature vector that will be fed to a fully connected layer for classification.

3.6.3 WDT CONV

This model employs voxel-level input fusion. It exploits Discrete Wavelet Transforms (DWT) to fuse MRI and radiation dose maps and a final 3D CNN backbone to extract relevant features.

DWT allows to decompose signals into low-frequency (approximation) and high-frequency (detail) components while preserving spatial (or temporal) resolution. Such components are called wavelets, which are small finite-duration oscillating waveforms

with an average value of zero. The decomposition is performed through a low-pass filter (L) and a high-pass filter (H) along each axis: L captures approximate information, while H captures detail information. It is an iterative process: initially the signal is fed to the two filters. Then the process can be iterated on the low-frequency output for further decomposition. In the case of a 3D image, which can be treated as a 3D signal, for each level decomposition we get eight sub-bands corresponding to low and high frequencies in all three dimensions:

- LLL: low frequencies along all dimensions
- LLH, LHL, HLL, LHH, HLH, HHL: mixed frequencies
- HHH: high frequencies along all dimensions

The lower frequencies embed large scale features of the image, while detail coefficients capture finer details, such as edges and textures. We visually investigated a potential way to fuse the coefficients extracted from the MRI and radiation dose maps. A visually promising result was obtained by averaging approximation coefficients and keeping detail coefficient with highest energy:

Approximation coefficient

$$\text{fused}_{\text{LLL}} = \alpha \cdot \text{coeffs\_MR}_{\text{LLL}} + (1 - \alpha) \cdot \text{coeffs\_RTD}_{\text{LLL}},$$

where $\alpha = 0.5$.

Detail Coefficients

$$E_{\text{MR},k} = \sum_{i,j,l} |\text{coeffs}_{\text{MR},k}(i,j,l)|^2, \quad E_{\text{RTD},k} = \sum_{i,j,l} |\text{coeffs}_{\text{RTD},k}(i,j,l)|^2,$$

$$\text{fused}_k = \begin{cases} \text{coeffs}_{\text{MR},k} & \text{if } E_{\text{MR},k} > E_{\text{RTD},k} \\ \text{coeffs}_{\text{RTD},k} & \text{otherwise.} \end{cases}$$

Then the fused imaged is obtained by applying the Inverse Discrete Wavelet Transform to the resulting coefficients.

The following figures 3.8 show the final result of the proposed fusion:

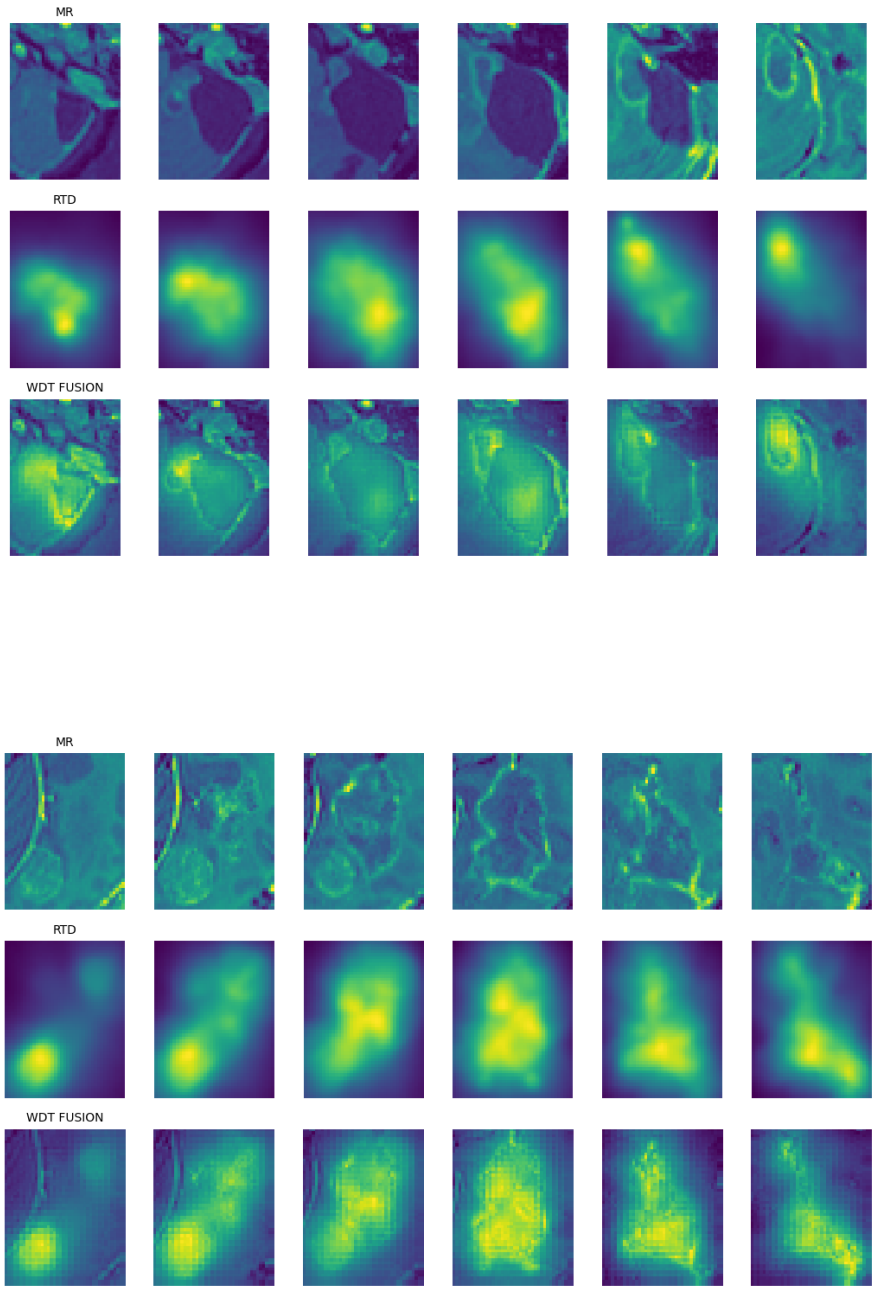

Figure 3.8: MRI and radiation dose maps fusion results through WDT.

The fusion resembles an "highlighting" effect upon the lesion, underling the regions with higher concentrations of radiation doses.

The 3D CNN backbone, which follows the line of the one adopted in the base model, consists of four sequential convolutional blocks, each one containing one 3D convolutional layer, one max pooling layer for spatial downsampling, one batch normalization layer, and ReLU activation applied to the convolutional outputs. The model increases progressively the number of feature channels from 1 to 256 across the convolutional layers. Finally, a global average pooling layer reduces the spatial dimension resulting in a compact feature vector of 256. Then, two fully connected layers are employed to project the extracted 256 feature to a 512 dimensional space, dropout is employed in between to reduce overfitting and the final fully connected layer re-compacts the feature vector back to the original dimensionality of 256.

Clinical data embedding is obtained through the MLP CD model and the resulting feature vector is concatenated to the to the one extracted from the 3D CNN backbone. A final fully connected layer computes the final prediction.

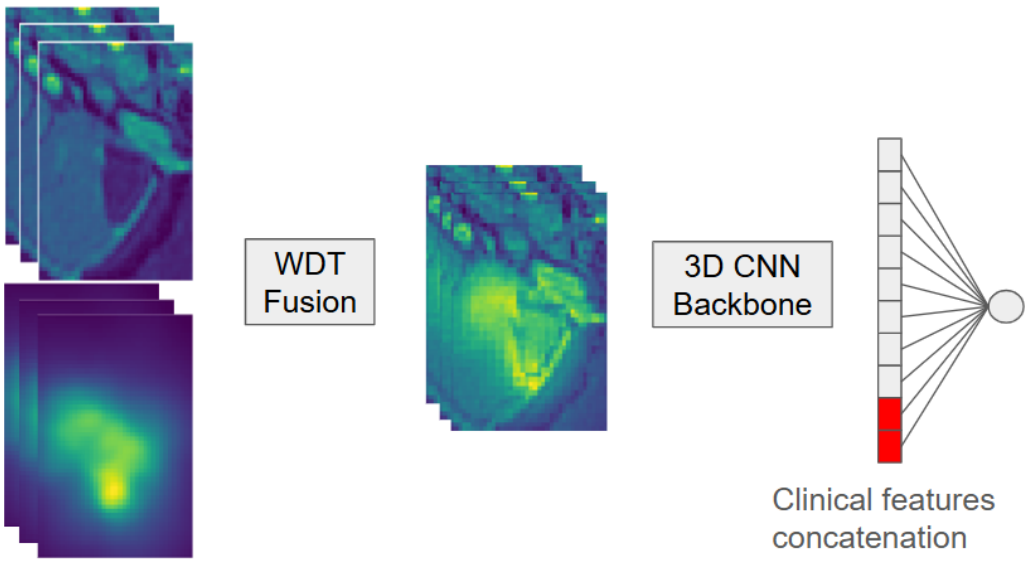

Figure 3.9: WDT CONV model architecture.

3.6.4 CONV LSTM

Jalalifar et al. [13] propose a deep learning solution for predicting the outcome of SRT, exploiting a InceptionResNetV2 backbone to extract local features from T1w and T2-FLAIR MRI slices and LSTM, Seq2Seq or transformer network to incorporate spatial dependencies between slices. Although our imaging includes T1w MRI and planned radiation dose maps, we tried to adapt this solution into our domain, as shown in fig. 3.10.

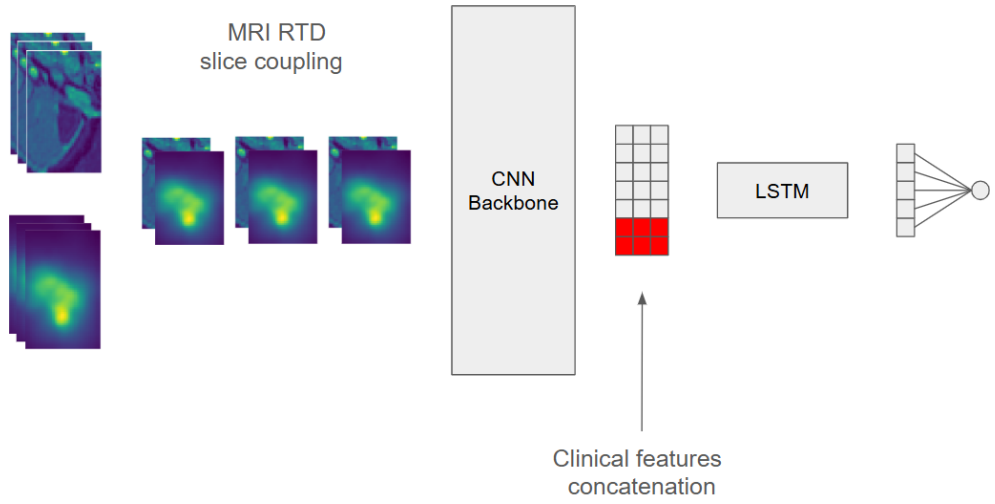

Figure 3.10: CONV LSTM architecture.

Initially 2D slices are extracted from both MRI and radiation dose maps. Then 2-channel slices are created by concatenating along the channel dimension the extracted 2D slices. The chosen backbone for feature extraction is ResNet18 in our case, which is less complex then the proposed InceptionResNetV2 network. The output of the CNN backbone results in a list of feature vectors encoding local spatial dependencies between MRI and radiation dose map slices.

Another reason why we chose to employ this solution is also because of the exploitation of clinical records. The authors collected different clinical information such as: histology, tumor location between (infratentorial or supratentoria), tumor size (longest diameter in mm), number of brain metastases, total dose (Gy), previous Whole Brain Radiation Therapy (WBRT) (yes/no), prior SRT/SRS (yes/no), GPA (from 0 to 4), age and gender. The clinical records are embedded through a MLP with one hidden size of 10. After an exhaustive search the final chosen records were histology, tumor location, tumor size, and number of brain metastases which have proven to be the most discriminant.

As we previously said, we decided to employ tumor size and number of brain metastases too which were not specifically provided in our dataset but were retrieved from the available clinical records. The authors fuse clinical features, extracted from the MLP backbone, and the features extracted by the imaging by concatenation. The sequence is then fed to a LSTM network with 2 layers and hidden size of 45. A final fully connected layer outputs the final prediction. In the article, the LSTM configuration with clinical features outperformed all the other models.

3.6.5 TransMed

Dai et al. [5] propose a multi-modal attention-based fusion for medical imaging classification. A CNN backbone and a transformer work together: the CNN manages to extract local dependencies whereas the transformer establishes long-range spatial dependencies between different modalities.

Transformers

Transformers, introduced by the article "Attention is all you need" [17], are a class of deep learning architectures, originally developed in the field of natural language processing (NLP). Nevertheless its applications have expanded to other tasks such speech processing and computer vision.

Self-attention mechanism

The key concept introduced by transformers is the so called Self-Attention mechanism. It allows to weigh the relevance of each input element to every other element in a sequence, regardless of their position allowing to capture long-range dependencies more effectively. Input vectors are linearly transformed into three different vectors: Q (query), K (key), and V (value) vectors.

$$Q = XW_Q, K = XW_K, V = XW_V$$

W_Q, W_K, and W_V are learned weight matrices.

We can think of Q as the target element for which we are trying to find relevant dependencies with other elements in the sequence. K, instead, represents the elements in the sequence from which we are trying to find relevant information with respect to Q. Finally, V contains the actual information of the elements in the sequence.

The attention weights are computed by applying the softmax function to the dot product between the Q vector with all K vectors, scaled by the square root of the dimension of the K vectors:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

The dot product between Q and K measures how much attention Q should pay to each key. Scaling by the square root of the dimension of the key vectors helps to stabilize gradients when the dimension becomes too large. The softmax normalization converts the scores into probabilities (sum equal to 1) producing the final attention weights. The attention weights are then used to weight the values (V), aggregating the information the query focuses on.

Multi-Head Self-Attention

Multi-Head Self-Attention is the component the allows the transformer to focus on different parts of the input sequence simultaneously, leading to richer and more diverse

representations. Instead of using a single attention mechanisms, multiple parallel attention heads are used. Each head performs scaled dot-product attention independently which are concatenated and passed through a final linear layer with a learnable weight matrix to produce the final output.

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_n)W_O$$

Where:

$$\text{head}_i = \text{Attention}(QW_{Q_i}, KW_{K_i}, VW_{V_i})$$

DeiT

Data-Efficient Image Transformer (DeiT) is a transformer model designed for image classification. It aims to achieve high performances, even with limited training data which suits perfectly to medical datasets suffering from data scarcity. DeiT follows the general transformer-based architecture: the input image is split into fixed-size patches and flattened into sequences. It introduces a distillation token to incorporate knowledge from a teacher model, like a CNN, into the training process. Within TransMed, the distillation token is omitted, and the extracted patches are fed through a ResNet network to extract local features. To the patch embedding is appended one class token and are added positional embeddings: the class token acts as a representative summary of the input, learning to aggregate information from al patches during training, whereas the positional embeddings encode the spatial relationships between patches. Positional embedding are needed since transformer don't inherently know the order of the input data. This is the final structure of a sequence of N patch embeddings of size D:

$$Z_0 = [z_{class} + e_0; z_1 + e_1; z_2 + e_2; ...; z_N + e_N]$$

$$\text{Shapes} = \begin{cases} z_{class} & [1, D] \\ e & [N + 1, D] \end{cases}$$

Where z_i is the patch embedding, z_{class} is the class token, and e is the positional embedding sharing the same size of the patch embedding. The embedded sequences are then fed to the multi-head self-attention layer. The final output is finally passed through a MLP consisting of linear layers separated by a GeLu activation, enabling the network to model complex relationships. Both multi head attention and MLP exploit skip connections, specifically:

$$z' = z + \text{MultiHeadAttention}(\text{LayerNorm}(z))$$

$$z'' = z' + \text{MLP}(\text{LayerNorm}(z'))$$

By defining as transformer encoder block the union of MHA and MLP, the final structure of the DeiT consists in many sequential layers of such blocks (depth). A final fully connected layer outputs the final prediction.

TransMed Architecture

We can split the TransMed architecture into 3 parts, as shown if Fig. 3.11:

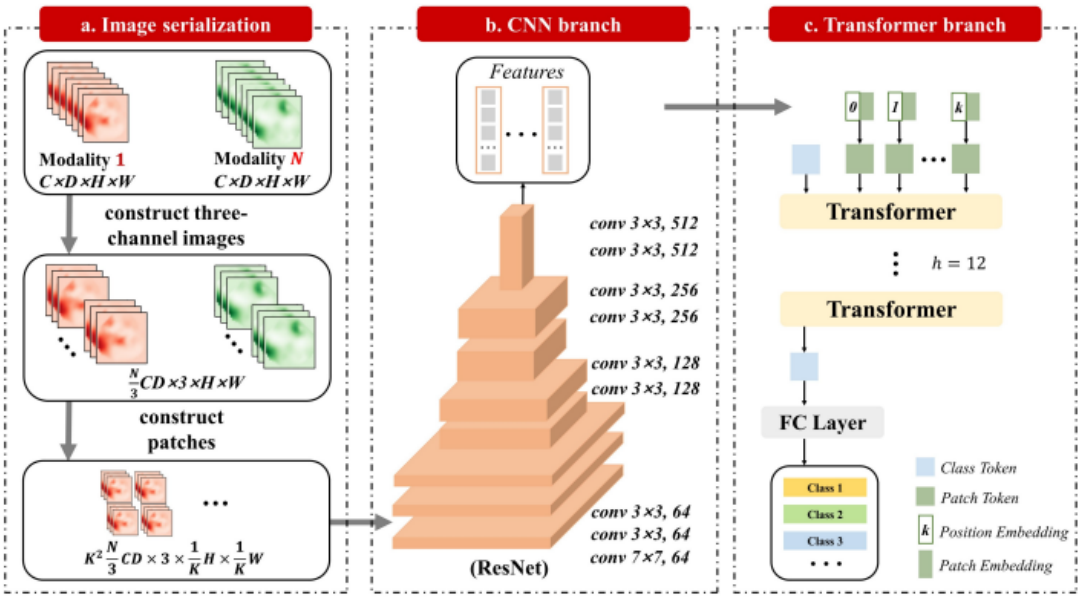

Figure 3.11: TRANS MED architecture. [5]

The first stage is in charge of creating the patch sequence: firstly the modalities are concatenated through the modality dimension, then 2D slice are extracted from the 3D imaging (the modality dimension is N=2 in our case, referring to MRI scan and radiation dose map). Then slices are grouped into triplets, creating 3-channel 2D slices. Finally fixed size patches are created: the patch size is an hyper-parameter, but as showed in the article, higher patch sizes tend to decrease the performances. The second stage consists into extracting features from the patch sequences: for our implementation we chose a ResNet18 backbone. After appending the class token and adding positional embeddings to the patches we conclude with the last stage where sequence patch embeddings are fed to the DeiT-based transformer model. The DeiT implementation was taken from the online github repository [18] of Francesco Saverio and adapted to the task.

3.7 Thresholding

At the end of the learning process, final logits are fed to a sigmoid which bounds the input within 0 and 1. Given the imbalanced nature of the data, the default threshold $t = 0.5$ might not be the optimal one. The optimal threshold depends highly on the task, since there may be applications where false positives and false negatives have different costs. In this work, we consider positive samples the recurrent lesions and negative samples the stable lesions. Recall, in treatment outcome prediction, might be considered more valuable, since false positives can lead to delayed interventions, worsening the prognosis.

Then, we chose to select the threshold leading to the maximum F1-score, which delivers the best compromise between precision and recall. During the training phase the threshold is computed on both training and validation set, and finally the two are averaged in order to deliver a more generalized one.

During cross-validation maintaining the fixed test set, at each iteration we store the obtained threshold and compute a global final average to be finally used when averaging the prediction obtained from the models trained on different folds.

Chapter 4

Results

4.1 Metrics

In this work we considered the following metrics:

1. **Accuracy**: proportion of correctly classified samples out of the total number of samples
2. **Precision**: proportion of true positive predictions out of all positive predictions
3. **Recall**: proportion of actual positives that the model correctly identified as positive
4. **Specificity**: proportion of actual negatives that the model correctly identified as negative
5. **F1 score**: harmonic mean of precision and recall
6. **Area under Receiver Operating Characteristic (AUROC)**: evaluates the ability of the model to distinguish between positive and negative classes at different threshold levels
7. **Area Under Precision Recall (AUPR)**: evaluates the trade-off between precision and recall at various thresholds

We report AUPR in addition of AUROC since it can give a more informative overview of the performances of model given the imbalanced data context.

4.2 Cross validation over the whole dataset

In order to exploit the maximum number of samples during training time, cross validation has been used upon the whole dataset. The number of chosen folds is 10, and as already explained in the methods section, the samples have been split in such a way that a subject cannot be part of more than one set. At each iteration, 9 folds are used for training and 1 is used for testing. An enough number of samples has been held out at each iteration

to constitute a validation set, respecting a similar overall ratio of recurrent and stable lesions..

For what concerns the training hyper-parameters, we chose the Adam optimizer and the reduce on plateau scheduler. This scheduler decreases the learning rate value by a factor of 0.1 after 5 epochs without any improvement in the validation loss. The maximum number of training epochs is 100, but early stopping is employed if no improvement of the validation loss is recorded of at least $1e-5$ after 10 epochs. Moreover, a batch size equal to 32 was used for all models. Regarding the focal loss hyper-parameters, α was set to 0.2 and γ was always tested between 2 and 3.

The MLP CD model was trained using a learning rate of $1e-3$, 0.3 dropout probability and $1e-4$ weight decay.

The BASE MODEL was trained using a learning rate equal to $1e-5$, dropout probability 0.3, weight decay $1e-4$ and γ equal to 2. The best performing configuration, reported in the table, makes use of clinical data.

The WDT CONV model was trained again using a learning rate equal to $1e-5$, dropout probability 0.1, weight decay $1e-4$ and γ equal to 2. This configuration exploited clinical data too.

The CONV LSTM model was trained using a learning rate equal to $1e-5$, dropout probability 0.1, weight decay $1e-4$ and γ equal to 3. For what concerns the LSTM network, the used configuration was 2 layers with an hidden size of 48. Still the results shown are obtained by exploiting clinical data.

Lastly, the TRANS MED model was trained with a learning rate of $1e-5$, dropout probability 0.2, weight decay $1e-4$, and γ equal to 2. The DeiT backbone was set with a depth attention of 8 and patch size equal to 1. This configuration does not make use of clinical data.

Table 4.1 shows the obtained results:

Model	ACC	REC	SPEC	PREC	F1	AUPR	AUROC
MLP CD	0.79	0.35	0.83	0.18	0.24	0.15	0.64
BASE MODEL	0.88	0.04	0.97	0.13	0.06	0.17	0.68
WDT CONV	0.86	0.22	0.93	0.25	0.23	0.18	0.60
CONV LSTM	0.85	0.22	0.92	0.22	0.22	0.16	0.65
TRANS MED	0.64	0.35	0.67	0.10	0.16	0.15	0.65

Table 4.1: Cross validation result on the whole dataset.

The MLP CD model shows minimum capabilities of discrimination between recurrent and stable sample: the AUPR is greater than the lower bound, represented by the proportion of the minority class, and the AUROC is greater than 0.5.

The other models did not achieve much greater results. By referring to the AUPR metric, BASE MODEL and WDT CONV reached the highest values of respectively 0.17 and 0.18. Additionally, BASE MODEL shows that the chosen threshold is not optimal given the low F1 score.

For what concerns the TRANS MED configuration, the best results were not obtained by exploiting clinical data. This could mean that the technique used to fuse clinical

features may not have been the best for this model. Moreover, the DeiT backbone should be able to deliver high performances thanks to the patch mechanism too: positional embeddings are able to decode the original position of the patch, demonstrating to have learned intrinsic characteristics of such image. Choosing a patch size equal to 1 means that a single patch covers the whole 2D slice: this choice was taken since the majority of the lesions occupy few pixels in many slices. Dividing them in smaller patches would lead to mostly empty slices from which extracting any kind of meaningful feature would be impossible.

Overall the obtained results demonstrate that the proposed models are able to extract minimum relevant relationships between clinical data, imaging and treatment outcome but clearly do not reach satisfactory results.

4.3 Cross validation with fixed test set

We chose to perform additional experiments by exploiting the same test set described in the article proposing the dataset [4]. In this way we can have a direct comparison with the baseline evaluations disclosed by the article (only F1 score, Precision, Recall, Specificity and Accuracy) and register additional performance metrics that can help us diagnostic the previously obtained low performances.

We ran the previously discussed configurations and registered the average performances, along with corresponding minimum and maximum values. We took advantage of the fixed external test set to choose a final optimal threshold by averaging the optimal thresholds chose within different training/validation splits, hoping to achieve an even more optimal and general one.

Table 4.2 reports the obtained results:

Model	ACC	REC	SPEC	PREC	F1	AUPR	AUROC
MLP	0.65	0.52	0.667	0.141	0.221	0.19	0.634
CD	(0.571, 0.747)	(0.0, 0.8)	(0.543, 0.84)	(0.0, 0.194)	(0.0, 0.304)	(0.088, 0.263)	(0.421, 0.772)
BASE	0.815	0.24	0.886	0.137	0.159	0.191	0.65
MODEL	(0.67, 0.89)	(0.0, 0.6)	(0.679, 1.0)	(0.0, 0.286)	(0.0, 0.286)	(0.113, 0.288)	(0.477, 0.753)
WDT	0.842	0.38	0.899	0.309	0.325	0.313	0.725
CONV	(0.802, 0.868)	(0.1, 0.6)	(0.827, 0.963)	(0.25, 0.385)	(0.143, 0.435)	(0.109, 0.485)	(0.368, 0.859)
CONV	0.793	0.24	0.862	0.167	0.175	0.206	0.676
LSTM	(0.626, 0.879)	(0.0, 0.6)	(0.63, 0.988)	(0.0, 0.333)	(0.0, 0.364)	(0.128, 0.326)	(0.504, 0.859)
TRANS	0.732	0.2	0.798	0.022	0.04	0.211	0.682
MED	(0.11, 0.89)	(0.0, 1.0)	(0.0, 1.0)	(0.0, 0.11)	(0.0, 0.198)	(0.134, 0.3)	(0.59, 0.791)

Table 4.2: Averaged performances reached between all folds. Minimum and maximum values are shown within brackets.

The model that reached the highest average AUPR is WDT CONV, with a value of 0.313.

Interestingly, by focusing on the minimum and maximum registered values we can see that some metrics can vary a lot. This could mean that some folds have much different sample distribution and the models are not able to generalize well. At different iterations the models try to gather discriminative features from the samples of the training set, which probably differ considerably from the ones of validation and, most definitely, from

the ones of the test set. Even though regularization techniques have been adopted, such as data augmentation, dropout, and weight decay, trained models still do not show enough generalization capabilities due to the high variability between samples of different sets.

4.4 Averaging predictions

A straightforward approach to leverage the knowledge from models trained on different cross-validation splits is to average their output predictions. Table 4.3 shows the performances reached with this method:

Model	ACC	REC	SPEC	PREC	F1	AUPR	AUROC
Baseline	0.90	0.10	0.89	1	0.18	-	-
MLP CD	0.65	0.50	0.67	0.16	0.24	0.22	0.70
BASE MODEL	0.89	0.40	0.95	0.50	0.44	0.37	0.75
WDT CONV	0.84	0.40	0.89	0.31	0.35	0.39	0.83
CONV LSTM	0.86	0.10	0.95	0.20	0.13	0.32	0.87
TRANS MED	0.89	0	1	0	0	0.33	0.79

Table 4.3: Results obtained by averaging prediction of models trained on different cross-validation splits.

By focusing mainly on AUPR and AUROC we can notice that WDT CONV reached the highest scores with 0.386 AUPR and 0.828 AUROC. CONV LSTM and TRANS MED, instead, show worse performances with respect to BASE MODEL by if we consider the AUPR metric. This time TRANS MED shows an F1 score equal to 0 which demonstrates again that the chosen threshold is not optimal.

These results indicate that it may be beneficial to actuate strategies that take into consideration the decision taken by the same model trained on different data. We should also point out that the slightly higher overall performances reached by choosing this test set might indicate a more fortunate data split. Nevertheless, the results registered in the last two experiments show valuable diagnostics that will help the development of future strategies.

4.5 Discussion

The results obtained by cross validating the whole dataset showed some minimum signs of learning, but are certainly not satisfactory. Additional experiments, by maintaining a fixed test set and cross validating the remaining training set, showed high variance within the performance metrics that can demonstrate the lack of generalization capabilities of the models when facing different sample distribution between sets.

The adopted threshold strategy did not always deliver the most optimal one. This can bee seen by looking at the minimums of recall and precision, which many times reached even zero. Moreover, when averaging the predicted probabilities, while maintaining the external test set, the chosen threshold for the TRANS MED model delivered 0 for both

precision and recall. This can be another sign of the high degree of difference of sample distribution between sets.

The MLP CD model, used as backbone for feature extraction of the clinical data, did deliver some minimal performances when used alone to predict treatment outcome. Since the imaging of small lesions do not contain much information, more discriminant clinical data could really help to take better predictions. Moreover, 3 out of 4 models benefited from employing clinical features extracted from this model which shows the importance of clinical data inclusion in strategies dealing with treatment outcome prediction.

As we have previously seen, the dataset suffers from a severe data imbalance where only 23 lesions are labeled as recurrent with respect to the 221 labeled as stable. Not only that, the majority of the imaging do contain few information: by looking at the masks shapes after resampling it is possible to notice that the majority of the lesions have a volume of around 10x10x10 voxels. Fewer of them have higher volumes, reaching a maximum of around 60x60x60. Since these models need fixed sized inputs, padding must be introduced to bring all the images to the same shape. In this work we chose to introduce constant padding, with 0 as constant value. This introduces artificial edges which the model is going to learn anyway. One possible solution may be to resize lesions with higher volumes, but this will lead to a considerable loss of valuable information. One could think also to consider regions that are wider than the mask of smaller regions in order to avoid padding. Still, this introduces information that do not directly impact the task of treatment outcome since the farthest is the considered region from the treated lesion the less it can contribute meaningfully to detect wether such lesion is recurrent or not after radiosurgery. Moreover, by increasing the considered regions of smaller lesions, we may risk to include other lesions and relative radiation treatment, affecting the prediction of the actual target lesion.

Chapter 5

Future works

Even though we could not reach satisfactory results, this thesis can deliver a valuable starting point for future works that aim to improve local tumor recurrence prediction after SRS treatment. These following potential extensions aim to build upon the limitations and findings of this thesis, paving the way for more robust methodologies.

For what concerns the raw MRI imaging, additional preprocessing techniques can be explored, which are widely known to directly affect models' performances. For example, skull stripping is able to remove the pixel intensities regarding the skull, allowing normalization steps to act exclusively on the target tissues. Additionally, more advanced techniques such as histogram intensity equalization and de-noising might be able to increase the quality of the raw MRI scans. In this work, we isolated the lesions and the relative radiation dose maps through the ROI masks and applied 0 padding to achieve the same dimension across different samples. The gathering of MRI scans with higher resolution may allow to zoom in smaller lesions, thus avoiding padding and loosing too much information.

Input fusion, in this domain, may be able to produce very good results as also shown by our experiments. More advanced techniques could manage to produce fused input representation capable of expressing the most relevant intel from both modalities. Despite requiring higher level of expertise, these strategies do manage to make the most out of the similar spatiality and structure, which can be obtained through proper resampling.

Planned radiation dose maps can surely be beneficial for what concerns SRS treatment outcome prediction since knowing how much radiation dose is delivered on each part of the lesion is a valuable information. Nevertheless, the current state of the surrounding tissues as a result of the tumor proliferation has been shown to be very valuable for this task. Additional imaging, among T2W, FLAIR, DWI, CT, and PET, may add important complementary information regarding, not only the lesion characteristics, but also its impact on surrounding healthy tissues and brain functionalities.

Yet, more information needs to be available in order to allow models to learn enough intrinsic features to successfully predict local progression after SRS. Deep learning models are known to be data-hungry, but in this case we really believe to be necessary: scarcity of the minority class along with the 1:10 imbalance really do hurt the final performances.

Additionally, we need to consider that a lot of the lesions are very small and, as a consequence, models relating to imaging may not be able to extract relevant information. Transfer learning should absolutely be taken into consideration: pretrained CNN backbones on larger brain MRI datasets might be able to effectively extract discriminant features from the current lesions, despite the limited number of samples.

We must underline the importance of complementary modalities: complementary in the sense that each one has the potential to compensate for the lack of discriminant characteristics of the others. Clinical data is surely beneficial in this sense, as shown by some related works and our results too. Given these considerations, future works may actuate output fusion techniques that are able to dynamically weigh the relevance of the final prediction between models relying on imaging and others relying on clinical features, by considering their level of confidence. Nevertheless, gathered clinical data must be carefully chosen: different studies demonstrate the relevance of GPA score, KPS index and number of metastases rather than usual demographics such as age and gender, for what concerns treatment outcome prediction. A more in-depth research needs to be performed in order to select preemptively known discriminative clinical data and eventually add and study the relevance on unexplored ones. LASSO regression is a popular strategy for what concerns feature selection as shown from related works. Lesion location, primary tumor location, and tumor histology need a standardized formatting in order to ensure consistency and free selection of specificity to successfully find the most discriminant settings for maximum exploitation of such features. A medical expert should supervise the processing of categorical values and suggest the choice of different levels of specificity which can be tested to find the most optimal one.

This work specifically deals with local progression of metastatic brain tumors. Nevertheless, treatment outcomes of GK treatment extend also to Pseudo Progression (PsP), temporary increase of volume as a response to the treatment, and Radio Necrosis (RN), increase of the contrast-enhanced area due to the death of healthy tissues. Prognosis of the described outcomes has potential relevance in the diagnostic domain, and can surely be impactful for treatment design and planning.

Chapter 6

Conclusions

In this thesis we addressed the challenge of local tumor recurrence prediction after Gamma Knife therapy. The dataset includes pre-treatment T1W MRI with gadolinium contrast, planned radiation dose maps, ROI masks and clinical data.

Preprocessing steps included imaging resample and registration in order to obtain aligned structures sharing the same spacing. Moreover, categorical clinical information have been standardized and generalized given the inhomogeneous terminology and the overly fine-grained values. Focal loss and data augmentation were adopted in order to mitigate the effects of data imbalance and scarcity of the minority class. Lastly, we tried to calculate an optimal classification threshold by choosing the one that maximized the F1 score during training time.

The proposed deep learning models adopt different multi-modal fusion strategies: BASE MODEL concatenates the features extracted from different modalities, WDT CONV fuses the imaging modalities voxel by voxel before the deep learning backbone, CONV LSTM fuses the extracted features from each modality through an LSTM network and finally TRANS MED exploits parallel self-attention mechanisms to find relevant relationships between the extracted features of different modalities.

Although showing minimal performances, the models did not reach satisfactory results: by cross validating the whole dataset the best performing models, by looking at the AUPR metric, where BASE MODEL and WDT CONV by reaching respectively 0.17 and 0.18 AUPR with correspondingly 0.60 and 0.68 AUROC. Every model, except from TRANS MED, benefited from the inclusion of clinical features which demonstrates their relevance. Additional experiments were conducted by using the same fixed test set proposed in the dataset's article and cross validating the remaining training set: the noticeable difference of minimum and maximum values of the registered metrics shows lack of generalization capabilities of the models when encountering high degree of difference of the sample distribution between sets. The highest average AUPR was reached by the WDT CONV model with a value of 0.313 with a relative average AUROC of 0.725. Finally, by averaging the predictions of the models trained on different folds the WDT CONV model reached the highest AUPR of 0.39 with an AUROC of 0.83, which demonstrates the potential of models jointly contributing to generate a final prediction.

Throughout this work, several challenges were encountered including domain-specific

imaging manipulation, data imbalance, scarcity of the minority class, and inhomogeneous clinical information. Despite these obstacles, minimal results were achieved, along with valuable insights that we hope will contribute to future works on this critical task. Improving performances in such a challenging context could encourage the development of better and richer datasets, ultimately magnifying the accuracy of local SRS outcome predictions and improving both treatment design and planning for patients with metastatic brain tumors.

Bibliography

[1] Case Western Reserve University. Mri basics, n.d. Accessed: November 25, 2024.

[2] Yihao Li, Mostafa El Habib Daho, Pierre-Henri Conze, Rachid Zeghlache, Hugo Le Boité, Ramin Tadayoni, Béatrice Cochener, Mathieu Lamard, and Gwenolé Quellec. A review of deep learning-based information fusion techniques for multimodal medical image classification. *Computers in Biology and Medicine*, 177:108635, Jul 2024.

[3] Ross Girshick Kaiming He Piotr Dollar Tsung-Yi Lin, Priya Goyal. Focal loss for dense object detection. *Facebook AI Research (FAIR)*, 2018.

[4] Yibin Wang, William Neil Duggar, David Michael Caballero, Toms Vengaloor Thomas, Neha Adari, Eswara Kumar Mundra, and Haifeng Wang. A brain mri dataset and baseline evaluations for tumor recurrence prediction after gamma knife radiotherapy. *Scientific Data*, 10(1):785, Nov 2023.

[5] Yin Dai, Yifan Gao, and Fayu Liu. Transmed: Transformers advance multi-modal medical image classification, 2021.

[6] AIOM. Associazione italiana oncologia medica, 2023. https://www.aiom.it/.

[7] Jan C. Buckner, Paul D. Brown, Brian P. O'Neill, Fredric B. Meyer, Cynthia J. Wetmore, and Joon H. Uhm. Central nervous system tumors. *Mayo Clinic Proceedings*, 82(10):1271–1286, Oct 2007.

[8] Animesh Saha, Sajal Kumar Ghosh, Chhaya Roy, Krishnangshu Bhanja Choudhury, Bikramjit Chakrabarty, and Ratan Sarkar. Demographic and clinical profile of patients with brain metastases: A retrospective study. *Asian J Neurosurg*, 8(3):157–161, July 2013.

[9] Nancy U Lin, Eudocia Q Lee, Hidefumi Aoyama, Igor J Barani, Daniel P Barboriak, Brigitta G Baumert, Martin Bendszus, Paul D Brown, D Ross Camidge, Susan M Chang, Janet Dancey, Elisabeth G E de Vries, Laurie E Gaspar, Gordon J Harris, F Stephen Hodi, Steven N Kalkanis, Mark E Linskey, David R Macdonald, Kim Margolin, Minesh P Mehta, David Schiff, Riccardo Soffietti, John H Suh, Martin J van den Bent, Michael A Vogelbaum, Patrick Y Wen, and Response Assessment in Neuro-Oncology (RANO) group. Response assessment criteria for brain metastases: proposal from the RANO group. *Lancet Oncol*, 16(6):e270–8, May 2015.

[10] Peng Du, Xiao Liu, Li Shen, Xuefan Wu, Jiawei Chen, Lang Chen, Aihong Cao, and Daoying Geng. Prediction of treatment response in patients with brain metastasis receiving stereotactic radiosurgery based on pre-treatment multimodal MRI radiomics and clinical risk factors: A machine learning model. *Front Oncol*, 13:1114194, March 2023.

[11] YU JIN CHA, WON IL JANG, MI-SOOK KIM, HYUNG JUN YOO, EUN KYUNG PAIK, HEE KYUNG JEONG, and SANG-MIN YOUN. Prediction of response to stereotactic radiosurgery for brain metastases using convolutional neural networks. *Anticancer Research*, 38(9):5437–5445, 2018.

[12] Se Jin Cho, Wonwoo Cho, Dongmin Choi, Gyuhyeon Sim, So Yeong Jeong, Yun Jung Baik, Sung Hyunand Bae, Byung Se Choi, Jae Hyoung Kim, Sooyoung Yoo, Jung Ho Han, Chae-Yong Kim, Jaegul Choo, and Leonard Sunwoo. Prediction of treatment response after stereotactic radiosurgery of brain metastasis using deep learning and radiomics on longitudinal mri data. *Scientific Reports*, 14(1):11085, May 2024.

[13] Seyed Ali Jalalifar, Hany Soliman, Arjun Sahgal, and Ali Sadeghi-Naini. Predicting the outcome of radiotherapy in brain metastasis by integrating the clinical and MRI-based deep learning features. *Med Phys*, 49(11):7167–7178, July 2022.

[14] V.P. Subramanyam Rallabandi and Krishnamoorthy Seetharaman. Deep learning-based classification of healthy aging controls, mild cognitive impairment and alzheimerâs disease using fusion of mri-pet imaging. *Biomedical Signal Processing and Control*, 80:104312, 2023.

[15] R. Beare, B. C. Lowekamp, and Z. Yaniv. Image segmentation, registration and characterization in r with simpleitk. *Journal of Statistical Software*, 86(8), 2018.

[16] Elham Karami, Hany Soliman, Mark Ruschin, Arjun Sahgal, Sten Myrehaug, Chia-Lin Tseng, Gregory J. Czarnota, Pejman Jabehdar-Maralani, Brige Chugh, Angus Lau, Greg J. Stanisz, and Ali Sadeghi-Naini. Quantitative mri biomarkers of stereotactic radiotherapy outcome in brain metastasis. *Scientific Reports*, 9(1):19830, Dec 2019.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000â6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[18] Francesco Saverio Zuppichini. Deit: Data-efficient image transformers. https://github.com/FrancescoSaverioZuppichini/DeiT, 2021.