Abstract

Due to the possible benefits introduced by Artificial Intelligence (AI) techniques, more and more efforts have been made for the development of non-invasive tools able to support clinicians in both diagnosis and prognosis, based on medical images. However, there are still uncovered issues which strongly limit their introduction into clinical practice, related to both medical data analysis and the implementation of such systems. On one hand, the unavoidable data variability, which characterizes multi-centre medical databases, due to both non- and biological reasons, heavily affects the achievement of robust, reproducible, and statistically relevant results. On the other, there are no clear recommendations that may support AI-based system implementation in the medical domain.

For these reasons, this PhD project aims to contribute to filling this gap focusing on the different decisional steps related to the analysis and management of a multi-centre medical image database, and the development of both Deep (DL)-based and Machine Learning (ML)-based models. Rectal cancer (RC) has been used as an exemplary oncological objective since it was possible to investigate several critical points.

The work is based on 1212 annotated pelvic Magnetic Resonance Images (MRIs), acquired by 12 different centres. Moreover, for a subgroup of patients (n=95) the tumour response to therapy was provided in terms of Tumour Regression Grade (TRG).

As the first step, I reviewed the state of the art related to the variability issue, assessing that even if there are different normalization approaches, none of them is

the preferred one, and also that a very limited number of studies have been conducted on real-world multi-centre MRI datasets. For this reason, I assessed the impact of such normalization methods on the variability across a database subgroup (n=88), considering different scenarios given by combining multiple vendor scanners and acquisition protocols. My project then focused on analysing the development pipeline steps of both DL and ML models, to automatically segment and predict the complete therapy response respectively.

Indeed, focusing on DL, I evaluated the impact on the performances of different image normalizations, networks' hyperparameters such as loss function and structure, and the training set by considering different construction approaches and sizes, assessing the importance of each decision. Similarly, focusing on ML, I evaluated the impact of extracting, normalizing and selecting radiomic features, combined with different models.

Thanks to these analyses, it has been possible to assess the crucial importance of normalizing the medical images in reducing the variability across the sequences and improving the performance of AI-based systems. Moreover, focusing on the development of DL-based systems, I assessed the impact of selecting the proper structure and loss function, according to the task and expected outcomes, and the best representative subgroup of samples to be used as training set through clustering methods. Finally, related to the ML-based predictive model, it has been assessed the complexity and the importance of defining the most suitable criteria for the feature selection according to the implemented algorithm.

Even if it has been not possible to define a unique pipeline generalizable to all multi-centre studies, this project presented several insights and methodologies that may contribute to raising interest, attention, and support toward a more conscious decision when implementing such systems.