Project Assignment: Cardiovascular disease diagnosis using cardiac magnetic resonance radiomics and machine learning

Introduction:

Cardiovascular diseases (CVD) persist as the leading cause of mortality, accounting for approximately one third of all deaths [1]. Their assessment is usually performed by means of cardiac magnetic resonance (CMR), the reference imaging modality for studying cardiac function and structure. Traditionally, few parameters, such as ejection fraction and volumes, are calculated from the CMR and evaluated by the medical professionals. Despite the importance of these traditional indexes, they fail to capture the full complexity of the cardiac tissue. For this reason, in recent years, radiomics has emerged as a novel image analysis technique to assess CMR [2, 3]. Radiomics involves extracting a high number of features that characterize the shape, intensity, and texture of the structures of interest.

Project Aim:

In this project, your aim is to leverage machine learning to automatically classify patients' examinations into five distinct classes using as predictors cardiac magnetic resonance radiomics.

Dataset Overview:

For the purposes of this project, you will utilize data extracted from the ACDC challenge dataset [4, 5]. The dataset contains 100 patients, evenly distributed across the five classes of interest:

1. Normal case (NORM),
2. Heart failure with infarction (MINF),
3. Dilated cardiomyopathy (DCM),
4. Hypertrophic cardiomyopathy (HCM),
5. Abnormal right ventricle (RV)

You are provided with the file ACDC_radiomics.csv which contains the radiomics for three structures of interest (left ventricle, myocardium, right ventricle) at two different time-points of the cardiac cycle (end-systole and end-diastole) extracted from the CMR and the respective segmentations of the structures of interest. The radiomics where calculated using the Pyradiomics library [3]. An example slice of a CMR and its respective segmentation used for calculating radiomics is provided in Figure 1. Additionally, patient height and weight data are included. Finally, the patients' class, i.e., the target output, is provided in the “class” column.

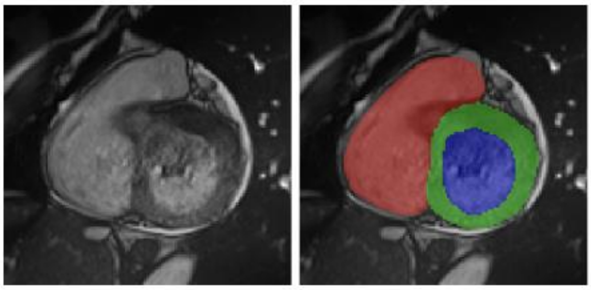

Figure 1: (a) Example 2D slice of a 3D CMR used for calculating the radiomics of the three structures of interest, (b) Segmentation of left ventricle (green), myocardium (blue) and right ventricle (red) overlaid on the CMR. Image reproduced from [1]

Datafile structure:

Each row of the file represents a patient, with columns corresponding to radiomic features, patient weight, height, and class labels. In total, you are provided with 643 radiomics of three different categories (107 per structure per time-point):

- Shape,
- First order, and
- Texture (GLCM, GLSZM, GLRLM, NGTDM, GLDM).

The naming convention for the characteristics in the .csv is as follows: *original_RadiomicCategory_FeatureName_Structure_Phase*, where *RadiomicCategory* represents shape, firstorder, glcm, glszm, glrlm, ngtdm, or gldm. *FeatureName* indicates the name of the feature. *Structure* refers to one of the structures of interest: myocardium (MYO), left ventricle (LV), or right ventricle (RV). *Phase* represents either end-diastole (ED) or end-systole (ES).

For example: original_ngtdm_Strength_MYO_ES.

For more details on the naming of the extracted radiomics, please refer to the Pyradiomics webpage[2].

Project tasks:

1. **Exploratory Data Analysis**: Conduct a thorough exploratory analysis of the dataset and comment on important findings that may influence model development. Please note that in this case, we suppose that the entire dataset (training, validation) is available.
2. **Training and validation set creation**: Divide the dataset into training and validation sets using a method of your choice, and justify your approach. Feel free to experiment with different splitting techniques (simple splitting, cross-validation).
3. **Baseline Model Development**: Utilize AutoML/H20 initially, followed by a tree-based classifier from scikit-learn, to develop a baseline model. Evaluate the model's performance using various metrics while discussing potential model issues, as well as the rationale behind the chosen evaluation metrics. Assess whether this model is acceptable, providing reasons for your conclusion. Computational difficulties may be encountered when performing AutoML due to the large number of predictive variables (features).

[2] https://pyradiomics.readthedocs.io/en/latest/features.html

4. **Feature Engineering and Model Enhancement**: Develop a pipeline that includes a dimensionality reduction or feature selection technique or a combination of both approaches as a prior step to classification. Repeat the experiments conducted in step 2. Try out at least two linear dimensionality reduction techniques, two nonlinear dimensionality reduction techniques, and three feature selection techniques. Provide details on the number of features used by each model, explaining the similarities and differences between the various approaches. Additionally, justify the concrete choices made in this process.
5. **Results Analysis**: Analyze the results to identify the most important features influencing model predictions. Compare and contrast the performance of different approaches. Finally, recommend the most suitable approach, if any, for clinical adoption. What steps do you think would be necessary to bring this model into real clinical use? Critically comment on which metrics should be used in this problem and why.

Project Guidelines:

1. The project should be developed in groups of 2 or 3 people. Please notify the instructors (amonleong@ub.edu, polyxeni.gkontra@ub.edu) about the composition of your group by **March 1st.**
2. You must submit a compressed file (.zip or .tar.gz) containing:
    a. A comprehensive report in PDF format, detailing the project workflow, choices, and findings.
    b. Python code implementing the project. Ensure to include all necessary files to run your code and replicate your findings.
3. The submission deadline for the project is **April 9th at 23:59**. Submit your project through the designated job on the Campus Virtual platform, named "ACDC Project" under the "Graded Activities" section. Please note that the job will close automatically after the deadline, and submissions will no longer be accepted.
4. Any instance of copying between groups or detected plagiarism will result in a zero grade for the project.
5. Sharing the ACDC_radiomics.csv outside of this course is prohibited without explicit permission from Polyxeni Gkontra.

Bibliography

[1] http://www.who.int/cardiovascular_diseases, Retrieved February 14, 2024.
[2] Gillies, R. J., Kinahan, P. E., & Hricak, H. (2016). Radiomics: Images are more than pictures, they are data. Radiology, 278(2), 563–577. https://doi.org/10.1148/radiol.2015151169
[3] Raisi-Estabragh, Z., Izquierdo, C., Campello, V. M., Martin-Isla, C., Jaggi, A., Harvey, N. C., Lekadir, K., & Petersen, S. E. (2020). Cardiac magnetic resonance radiomics: basic principles and clinical perspectives. European Heart Journal Cardiovascular Imaging, 21(4), 349–356. https://doi.org/10.1093/ehjci/jeaa028
[4] https://pyradiomics.readthedocs.io/en/latest/, Retrieved February 14, 2024.
[4] Bernard, O., Lalande, A., Zotti, C., et al. (2018). Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Transactions on Medical Imaging, 37(11), 2514–2525. https://doi.org/10.1109/tmi.2018.2837502