IMA205 repport on 2025 challenge

Cardiac Pathology Prediction

Quentin Revillon

Student at Télécom Paris, Polytechnique Institute of Paris

Pietro Gori, Loïc Le Folgoc

Abstract—In this report, we detail the methodology followed to perform cardiac pathology prediction on the ACDC dataset (Automatic Cardiac Diagnosis Challenge). The dataset contains data from 150 multi-equipment CMRI recordings with reference measurements, segmentation of the myocardium, left ventricle, and right ventricle, as well as disease classification from two medical experts. In the initial challenge, heart segmentation and patient classifications were missing for the last 50 patients. However, in this simplified version, only the segmentation of the left ventricle was missing. As a result, we performed left ventricle segmentation using morphological transformations on the myocardium and border detection to segment the left ventricle, which is located inside the myocardium. Cardiac parameters, such as ejection fraction and ventricle volumes, were then calculated from the segmentation masks. Furthermore, these parameters were used as features to train a random forest and an MLP. The dataset was split into training/validation (90% for cross-validation) and testing (10%). The most important features were extracted, and different classifiers were trained afterwards.

1. Introduction

Cardiac parameters such as left ventricular ejection fraction, volumes of the left and right ventricles, and myocardial thickness are routinely calculated to diagnose whether a subject is healthy or diseased. To compute these parameters, our first objective was to segment the left ventricle (the only missing segmentation) by filling the empty interior of a mask, combining the available left ventricle and myocardium segmentations. Afterwards, we computed classification parameters according to those used or suggested by Isensee et al. (2020), Khened et al. (2019), and Wolterink et al. (2016). The most relevant features were selected based on random forest importance. We then trained and fine-tuned a random forest, an MLP, an SVM, an XGBoost model, and a voting classifier on 90% of the dataset, and made predictions on the remaining 10%.

2. Left ventricle segmentation

2.1. main method

To develop the left ventricle segmentation method, we used the complete available segmentations from the first 100 patients without splitting the dataset into training and testing sets for this part. It was observed that the left ventricle was always located inside the myocardium in all slices. More precisely, it was consistently inside the (almost) closed shape formed by the right ventricle and the myocardium. Therefore, we created a mask by combining the right ventricle and myocardium masks, then filling the internal void.

The void was detected using OpenCV's `findContours` method, followed by filling the detected shape and computing the difference between the filled shape and the original mask. However, this method occasionally segmented a few pixels outside the myocardium, between the right ventricle and the myocardium, leading to two or more disconnected regions being segmented. We selected the largest connected component as the left ventricle.

To evaluate this segmentation method, we removed the left ventricle from all available segmentations and re-segmented it. Our goal was to achieve perfect segmentation compared to the available ground truth. To do this, we used an equality metric to compare our segmentation with the original one.

The method achieved perfect segmentation on all slices for 98% of the patients at diastole and 95% at systole.

The method was mainly failing when the myocardium was not a closed shape as in the first MRI slice of patient 11 (at the top of the myocardium).

To solve this problem, we performed a morphological dilation of

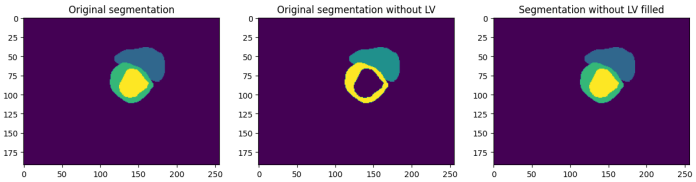

Figure 1. Successful segmentation of patient 11 's left ventricle in first slice

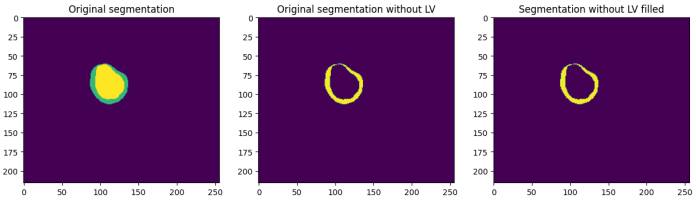

Figure 2. Failed segmentation of patient 11 's left ventricle in first slice

the myocardium using a cross-shaped structuring element. We then segmented the left ventricle as previously described. To compensate for the reduced area resulting from the myocardium dilation, we also applied dilation to the segmented left ventricle using the same structuring element. This method achieved near-perfect segmentation. However, a few isolated pixels between the segmented left ventricle and the myocardium remained unsegmented (after reinserting the segmented left ventricle into the myocardium mask). These pixels were iteratively added to the left ventricle. Although another iteration of the main segmentation method might have sufficed, this refinement led to perfect segmentation for all 100 patients at both systole and diastole, except for the first slice of patient 29. In that slice, the myocardium was too open for the method to work. Since this was the only failure case in the entire dataset, we chose to disregard it.

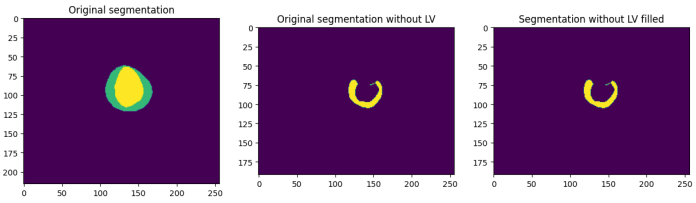

Figure 3. Successful segmentation of patient 11 's left ventricle in first slice

3. Diagnosis

3.1. Defining cardiac parameters

To develop an automatic heart disease diagnosis algorithm we extracted cardiac parameters from our previously segmented images. First we decide to exhaustively compute all the static cardiac parameters used or suggested in Isensee et al. (2020), khened et al (2019) and Wolterink et al. (2016). Indeed we only had access to MRI images at diastole and systole instant. Using these parameters, we computed a PCA of our training set and added the principal components explaining 95% of the parameters variation.

3.2. Selecting features to train ML algorithms

We selected our features by using random forest importance after training a large random forest of 100000 trees on the training part

of the dataset. To select the features we used SelectFromModel of sklearn.feature_selection, selecting all the features more important than the average level of importance. Between 14 and 20 features where selected, depending on the part of the dataset randomly selected for training. The number of trees used was sufficient to lead to constant feature selection for several trainings of the random forest classifier.

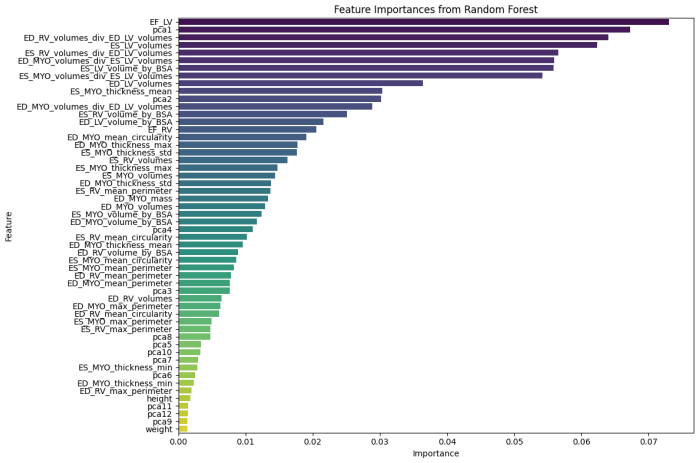

Figure 4. Successful segmentation of patient 11 's left ventricle in first slice

The parameters selected as features for our model are:

Table 1. Top 15 features automatically selected by Random Forest

Feature description
Left ventricular ejection fraction
1st principal component (PCA)
End-diastolic volume ratio RV/LV
Left ventricular end-systolic volume
Left ventricular end-diastolic volume
End-systolic volume ratio RV/LV
End-systolic myocardial volume to LV ratio
End-diastolic myocardial volume to LV ratio
Myocardial volume to LV ratio at end-systole (duplicate or derived feature)
Left ventricular end-systolic volume indexed to BSA
Mean myocardial thickness at end-systole
Left ventricular end-diastolic volume (duplicate or alternate definition)
Left ventricular end-diastolic volume indexed to BSA
2nd principal component (PCA)
Myocardial volume relative to LV end-diastolic volume

3.3. training of different models

We decided to train different models to try to use voting classification to build a more powerful model. In this aim, we decided to train a Random Forest, an MLP, an SVM, and an XGBoost, and a voting classifier combining all of them.

To train them and find optimal parameters we used cross-validation and GridSearch on 5 stratified folds to keep the same distribution of classes in each fold. After fine tuning the model we predicted the classes on testing dataset composed of 10 patients.

The models that seemed to perform the best were the Random Forest Classifier and the MLP. Thus we decided to run the whole pipeline, from random training/testing split, grid search of the optimal parameters to predictions on the testing data set, until we reached perfect accuracy on the testing dataset, giving us confidence in the ability of the model to generalize well.

The other models were trained with the same part of the dataset as the Random Forest and the MLP, leading to the following submission results:

Table 2. Models accuracy on the unclassified 50 last patients

Model	Parameters	Private submission accuracy
Random Forest	min samples leaf: 1 min samples split: 2 number of trees: 200	0.88
MLP	number of hidden layers: 1 number of neurons: 50 solver:Adam alpha L2 reg.: 0.0001 learning rate init:0.01	0.88
SVM	C: 10 gamma: scale kernel: linear	0.77
xgboost	colsample bytree: 0.7 learning rate: 0.01 max depth: 3 nb estimators: 100 subsample:0.5	0.77
voting model	voting: soft models: rf, mlp, xgboost max depth: 3	0.77

4. Conclusion

The method used for segmentation was highly efficient, achieving almost perfect accuracy on this simplified version of the initial challenge. We would have needed to use a different method if the issue of an overly open myocardium had occurred more than once in the entire dataset. Furthermore, we would have employed deep learning techniques if all the segmentations had been missing.

For classification, the small size of the dataset prevented us from using larger models, especially for the MLP. Additional features could have been computed, and noise could have been introduced to train larger models and improve generalization.