Télécom Paris

Classification of Cardiac Pathologies

Report for the IMA 205 Challenge 2025 – Kaggle Competition [2]

Romain Planchon

IMA 205 – Image Processing and Machine Learning Project
Instructors: Loic Le Folgoc and Pietro Gori
Academic Year 2024–2025

May 1, 2025

Contents

List of Figures

List of Tables

List of Abbreviations

CMRI	Cardiac Magnetic Resonance Imaging
mRMR	Minimum Redundancy Maximum Relevance
SVM	Support Vector Machine
MY	Myocardium
RV	Right Ventricle
LV	Left Ventricle
EF	Ejection Fraction
ES	End Systole
ED	End Diastole
GLRLM	Gray Level Run Length Matrix
GLCM	Gray Level Co-occurrence Matrix

Chapter 1

Introduction

1.1 Context and Motivation

Cardiac function assessment plays a central role in clinical cardiology, as it supports early diagnosis, risk stratification, and therapeutic decisions. Several cardiac diseases can remain asymptomatic in their early stages but may lead to serious complications such as heart failure or sudden cardiac arrest if not detected in time.

Over the past few years, various approaches for computer-aided diagnosis (CAD) based on cardiac magnetic resonance imaging (CMRI) have emerged, offering non-invasive, reproducible, and high-resolution insight into cardiac structure and function. Among the most studied benchmarks in this domain is the ACDC dataset, which has served as a reference for model development and comparison.

The Automated Cardiac Diagnosis Challenge (ACDC) has thus become a competitive and widely studied task. Its objective is to classify patients into different diagnostic categories using features extracted from CMRI data. Despite the diversity of existing methods, improving the classification score on this dataset remains a strong driver of innovation in cardiac image analysis.

1.2 Goal of the Challenge

The aim of the ACDC challenge is to predict the diagnostic class of each subject from their CMRI exam. The dataset is structured around five clinical categories:

- Class 0: Healthy controls
- Class 1: Myocardial infarction
- Class 2: Dilated cardiomyopathy
- Class 3: Hypertrophic cardiomyopathy
- Class 4: Abnormal right ventricle

1.3 Available Data

The full dataset consists of cardiac MRIs from 150 patients. Labels are provided for the first 100 cases, which constitute the training set. The remaining 50 form the test set and are used for performance evaluation.

Each subject's CMRI exam contains two timepoints in the cardiac cycle: end-diastole (ED) and end-systole (ES). For the training set, segmentation masks are available for three anatomical regions: the left ventricle (LV), the right ventricle (RV), and the myocardium (MY), on both timepoints.

In the test set, however, only partial segmentations are provided as the segmentation of the LV is missing.

Chapter 2

Segmentation of the Left Ventricle

Automatic segmentation of the left ventricle (LV) in CMRI images is a key step in the computation of features used for disease classification. Since the training set includes ground-truth segmentations of the LV, we can use this data to evaluate the performance of our segmentation method. To assess segmentation quality, we compute the Dice similarity coefficient between the predicted and reference masks:

$$\text{Dice}(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|}$$

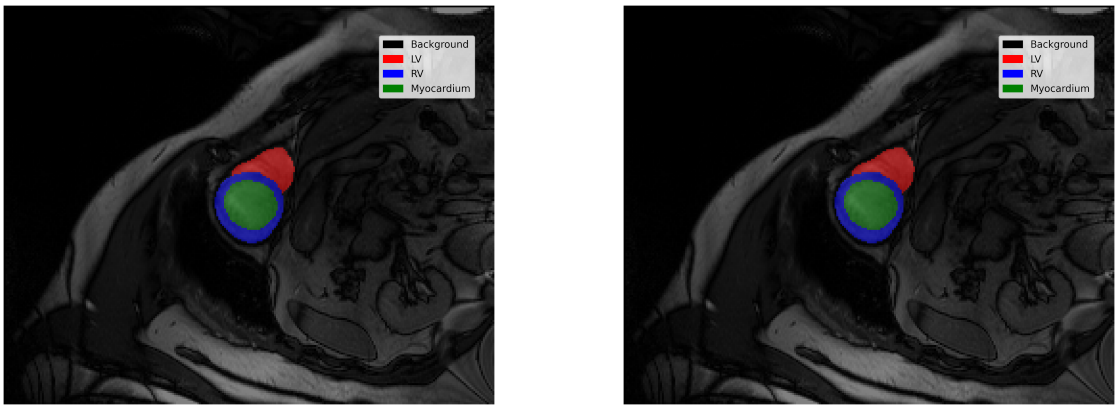
The Dice score ranges from 0 to 1, where a value of 1 corresponds to perfect overlap between the predicted segmentation and the ground truth, and a value of 0 indicates no overlap at all.

If our method achieves a satisfactory average Dice score on the training data, we will apply it to the test set in order to generate the missing LV segmentations.

2.1 General Case

2.1.1 Method

In more than 99% of slices (1895 slices over 1902), the LV cavity is completely enclosed within the myocardium. We first remove any existing LV label and convert the myocardium mask into a binary image. A standard *hole-filling* operation is applied to fill the central region inside the myocardium. All newly filled pixels are labeled as belonging to the LV. This method is fast and effective as long as the myocardium forms a closed contour.



(a) True segmentation

(b) Predicted segmentation

Figure 2.1: Comparison between true and predicted segmentation in the case where the myocardium is closed. (slice 5 from trainning 20 ES)

2.1.2 Performance

In the case were the myocardium forms a closed countour this method gets a average dice score of 1.This corresponds to a perfect score.

2.2 Handling Open Myocardium Cases

2.2.1 Method

However in some slices, the myocardium mask has small openings. This prevents the hole-filling operation from producing any result. In such cases, we apply a fallback procedure:

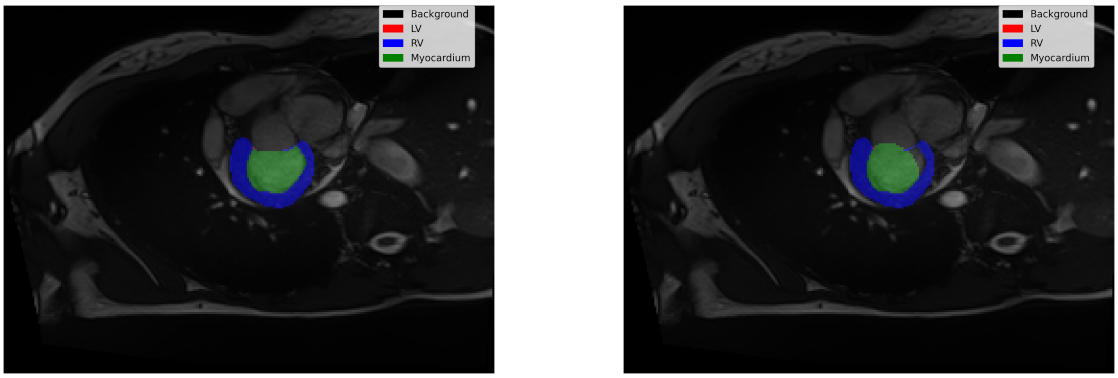
1. Compute the centroid of the myocardium pixels.
2. Extract all closed contours from the myocardium mask.
3. Estimate an average inner radius based on the distances from the centroid to the closest contour.
4. Perform a 4-connected region growing (flood fill) from the centroid, limited by the estimated radius and stopping at the myocardium boundary.

2.2.2 Performance

On the seven cases where the myocardium wasn't surrounding entirely the left-ventricle this technique got the following performances :

- Average Dice score: 0.958
- Maximum Dice score: 0.994
- Minimum Dice score: 0.896

This method is relatively naive, but as there are not many slices that required this adjustement I did not try more sophisticated methods that could have achieved better scores.



(a) True segmentation (b) Predicted segmentation

Figure 2.2: Comparison between true and predicted segmentation in the case where the myocardium isn't fully closed. The slice chosen was the one with the worst dice score. (slice 0 from trainning 29 ED)

2.3 Conclusion

Nonetheless, it should be noted that the method heavily relies on the quality of the prior myocardium segmentation. Errors or gaps in the myocardium mask could directly impact the final LV cavity segmentation. But apart from that, across all samples, our method achieves an outstanding average Dice score of 0.999, demonstrating its high robustness and accuracy for LV cavity reconstruction in CMRI images.

Chapter 3

Features

3.1 Feature Extraction

Now that we have acess to full segmentations of the left ventricle (LV), right ventricle (RV), and myocardium (MY), we utilize them to extract meaningful features.

In cardiac MRI applications such as pathology classification, shape features often capture the dominant differences between classes, making them highly predictive. Most existing solutions for this problem rely heavily on volume-related features.

Although such shape-based features perform well, relying solely on them means ignoring a large amount of potentially useful information. In deed pixel intensities and local variations in the regions of interest are totally set aside. Therefore, I chose in this challenge, to investigate not only shape related features but also on intensity and texture-based features as they could positively impact classification performance. I inspired my self from recent studies [4, 1] that demonstrated the potential of radiomics to extract complementary information from cardiac MRI. These studies employed the Python library `PyRadiomics`, which facilitates the extraction of a wide range of quantitative imaging features with minimal preprocessing effort.

Here are the categories of extracted features that I will use:

- **Shape features**: volume, surface area, sphericity, flatness, and others — for a total of 14 shape descriptors.
- **First-order intensity statistics**: mean, standard deviation, energy, entropy, etc.
- **Texture features from the Gray Level Co-occurrence Matrix (GLCM)**: autocorrelation, contrast, dissimilarity, homogeneity, maximum probability, etc.
- **Texture features from the Gray Level Run Length Matrix (GLRLM)**: short/long run emphasis, gray-level/run-length non-uniformity, and others.

Depending on whether we compute all these categories, the total number of features extracted ranges from approximately 100 to 400. Given the small size of our dataset, reducing this number is crucial to avoid overfitting.

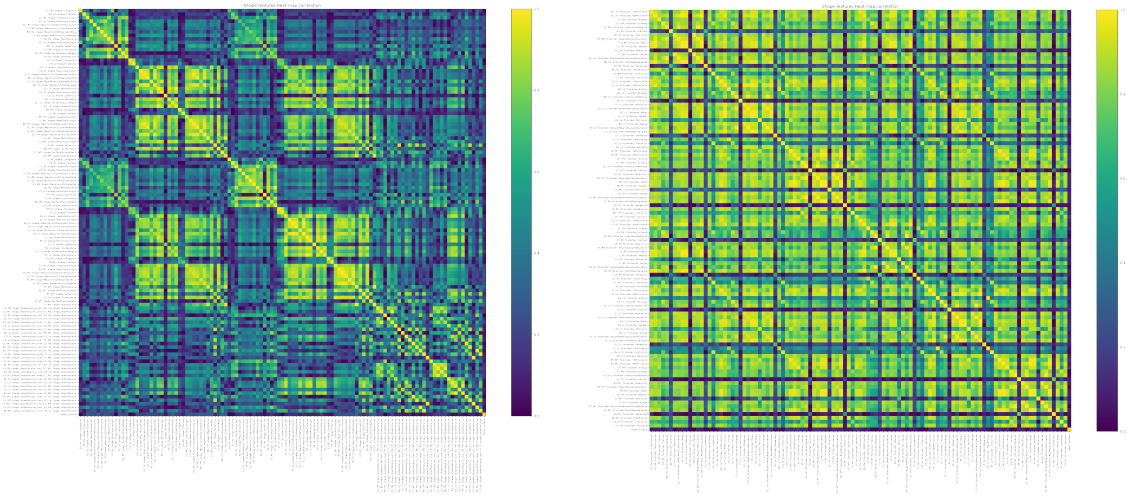
3.2 Feature Reduction

3.2.1 Variance and Correlation Filtering

The first step in reducing the feature space is to eliminate features with very low variance using a simple threshold. These features provide little to no discrimination between samples.

Next, we apply correlation filtering. Highly correlated features are often redundant and uniformative, so we remove one feature from each pair with a correlation above a fixed threshold. Performing this process revealed an interesting property: intensity and texture features tend to be much more correlated with each other than shape features are. We observe this phenomenon when visualizing the correlation matrices, which display distinct clusters of interdependent features.

We include the correlation heatmaps for shape and first-order features below; the others are provided in the appendix.



(a) Shape features (b) First-order intensity features

Figure 3.1: Correlation heatmaps of different feature types. Shape features (left) show relatively low interdependence (blue), while intensity features (right) form strong correlation cluster (yellow).

This observation helps to explain why shape features are often preferred in classification tasks and it is not the only factor. Unlike intensity-based features, shape features are independent of grayscale normalization, since they are derived directly from binary segmentation masks. In contrast, intensity-based features are sensitive to image contrast, scanner parameters, and require careful preprocessing to ensure comparability. Without proper normalization, their predictive power may be compromised.

3.2.2 Preprocessing Considerations

For any feature that relies on raw image intensities (i.e., anything but shape), it is crucial to apply standardized preprocessing steps such as resampling, bias correction, and intensity normalization. Neglecting these steps may introduce variance unrelated to the underlying anatomy or pathology.

3.2.3 Overfitting

Even after variance and correlation filtering, we obtain in our application a number of remaining features too high. A classic rule of thumb suggests limiting the number of features to approximately one per ten samples.

This constraint exists to mitigate the risk of overfitting, where a model performs well on training data but fails to generalize to new examples. Overfitting typically occurs when the

model captures noise or patterns specific to the training data, rather than true underlying relationships.

Chapter 4

Classification Model and Feature Selection

At this point, it is important to consider which model we will use for the final analysis. While Random Forest yields strong performance in the ACDC challenge litterature, when we want to take into account other features than shapes, it presents a limitation. Empirically when we allowed the RF model to freely select the most informative features, it consistently favors shape features over all others. When looking at the random forest feature performance of texture or signal intensities, they are always the last.

This behavior can be explained by the nature of the features themselves. Shape descriptors tend to be more stable, less redundant, and capture macro-level anatomical differences that are highly predictive, particularly in our context, where the disease causes significant anatomical and shape alterations. In contrast, intensity and texture features, though potentially informative, are more prone to noise and inter-correlation, reducing their relative importance in the model's decision-making process.

To be able to evaluate different method let's first compute a baseline model.

4.1 Baseline Model Performance

For our baseline model, we used only shape features with a Random Forest classifier.

4.1.1 Features used

After the initial feature extraction, we obtained *115* shape-related features. In this total I counted ; the standard PyRadiomics shape features, all possible voxel volume ratios as well as `body_surface_area` (BSA) computed from weight and height. Volume ratios are computed to capture relationships similar to the ejection fraction (EF), a metric repeatedly cited as highly informative for this classification problem. And BSA is added based on the intuition that larger body surface area correlates with larger cardiac volumes. After applying a feature selection pipeline combining a variance threshold and a correlation threshold as detailed before, we reduced the feature set to *57* features. For this baseline model, we retain all remaining features after this reduction.

4.1.2 Performance

Model evaluation was conducted using 5-fold cross-validation. Performance metrics used are summarized in Table 4.1.

Metrics Definitions

Testing Score: A weighted average of the private and public Kaggle scores:

$$\text{Testing Score} = 0.7 \times \text{Private Score} + 0.3 \times \text{Public Score}$$

Entropy Score: Measures uncertainty of predictions:

$$\mathcal{H}(\mathbf{p}) = -\sum_{i=1}^{K} p_i \log p_i$$

Entropy close to 0 means high certainty; close to 1 means high uncertainty.

Margin Score: Difference between the top two predicted class probabilities:

$$\text{Margin}(\mathbf{p}) = p_{\text{top1}} - p_{\text{top2}}$$

Higher margins indicate more confident predictions.

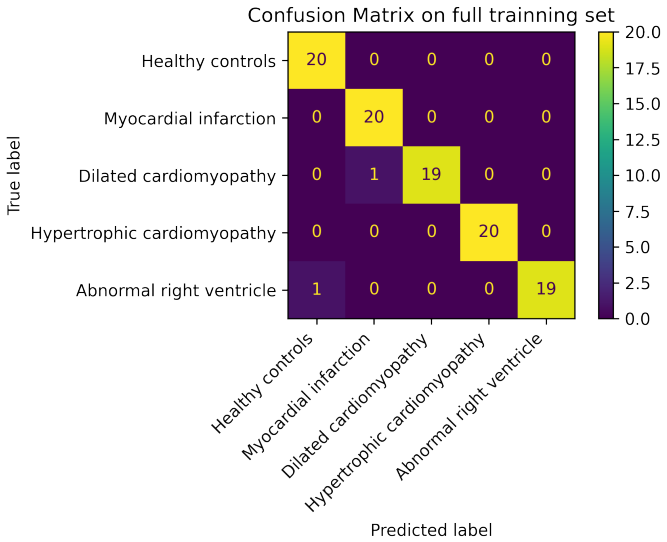


Figure 4.1: Confusion matrix for the baseline Random Forest model using only shape features.

Table 4.1: Performance metrics for the baseline Random Forest model.

Model	Train Score	Val. Score	Entropy	Margin	Testing Score
Baseline RF	0.995 ± 0.006	0.960 ± 0.037	0.696	0.757	0.895

These results set a strong initial benchmark. Indeed despite the model being a little

uncertain (relatively low margin), the model didn't overfitt as much as I thought it would. In the next sections, we will investigate whether alternative models or additional feature types (intensity, texture) can improve performance beyond this baseline.

4.2 Filtering Features

To improve generalisation, I applied a univariate ANOVA filter (SelectKBest with `f_classif`) to reduce the feature set from 57 to 15. The idea was to limit overfitting by selecting only the features whose values differed most between classes.

At this point, I did not have a clear expectation about performance: the reduction in feature count could help regularise the model, but it also risked discarding complementary information. This trade-off between simplifying the model and preserving predictive signal made the outcome uncertain. As the Table 4.2 below shows, in comparison with the baseline model, this filtering step led to a drop in accuracy.

In their foundational paper on feature selection, Guyon and Elisseeff [3] argue that the predictive power of a feature subset lies not in the individual strength of each variable, but in their collective synergy. Some variables may appear weak on their own yet become highly informative when combined with others. Motivated by this idea, I explored a more sophisticated filter: minimum-Redundancy Maximum-Relevance (mRMR), which retains features that are both informative and minimally correlated.

Surprisingly, performance decreased slightly compared to the ANOVA-based selection. While mRMR promotes diversity, it may have selected more varied but individually weaker features. In contrast, ANOVA likely retained more directly discriminative variables, even if partially redundant.

These experiments did not improve the model's performance, but they provided useful insight: aggressive pre-filtering is not always beneficial with Random Forests, which already manage redundancy through bootstrapping. Most importantly, they reinforce the idea that feature selection strategies should be evaluated empirically rather than assumed optimal.

Table 4.2: Performance of Random Forest after SelectKBest (k = 15).

Model	Train Score	Val. Score	Entropy	Margin	Testing Score
Random Forest (ANOVA)	0.993 ± 0.006	0.970 ± 0.040	0.792	0.817	0.874

Table 4.3: Performance of Random Forest after mRMR feature selection.

Model	Train Score	Val. Score	Entropy	Margin	Testing Score
Random Forest (mRMR)	0.995 ± 0.006	0.960 ± 0.037	0.740	0.779	0.857

These experiments did not improve the model's performance, but they provided useful insight: aggressive pre-filtering is not always beneficial with Random Forests, which already manage redundancy through bootstrapping and subspace sampling. Most importantly, they reinforce the idea that feature selection strategies should be validated empirically rather than assumed optimal.

At this stage, I believe that using only shape features, no matter how carefully preprocessed or selected, is unlikely to yield significantly better results. It is time to explore whether adding intensity and texture features can improve classification. However, with Random Forests, this is nearly impossible: even when these features are provided, the model tends to ignore them and build a shape-only solution. This is likely due to the model's internal variable selection mechanism, which favors features that produce strong splits early in the tree-building process. To address this issue, I turned to a different type of classifier, one that does not rely on greedy feature selection but instead considers all features together. I chose the Support Vector Machine (SVM), a model known for combining multiple moderately informative variables through margin maximisation. This approach has already been explored in the context of cardiac radiomics by Cetin et al. [1], and is particularly well suited to high-dimensional problems when features are properly scaled and decorrelated.

4.3 Support Vector Machine with Radiomics Features

To address the limitations of Random Forests and fully leverage the complementary information contained in radiomics features, I switched to a Support Vector Machine (SVM) classifier. Unlike Random Forests, which tend to ignore weaker features, SVMs optimise a global margin and can combine multiple moderately informative features when properly scaled and decorrelated.

Starting from a feature set of 40 variables selected using mRMR on shape, first-order intensity statistics, GLCM and GLRLM textures. I applied a Sequential Feature Selection (SFS) procedure. It iteratively adds the most useful features while monitoring validation performance. This process allowed me to reduce the feature set to 20 variables without loss in predictive power.

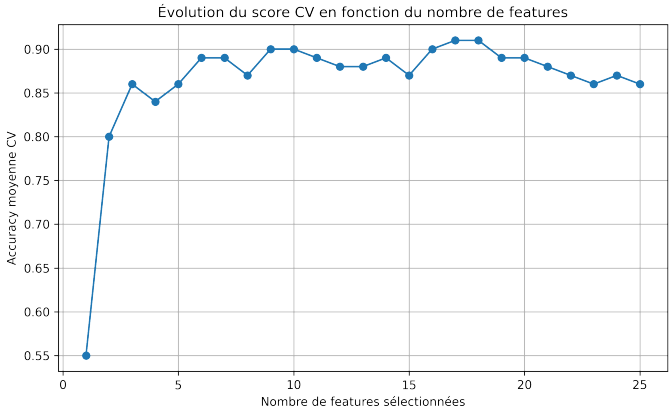


Figure 4.2: Accuracy evolution vs number of features used

Table 4.4: Performance of SVM trained on shape, intensity and texture features (20 features from SFS).

Model	Train Score	Val. Score	Entropy	Margin	Testing Score
SVM (SFS on mRMR)	0.970 ± 0.017	0.920 ± 0.081	0.639	0.739	0.895

The resulting model achieved the same testing score as the baseline Random Forest trained solely on shape features with fewer number of features (parameters). This suggests that, when

paired with an appropriate classifier, non-shape radiomics features can match the predictive performance of shape-based models. More importantly, it validates the potential of combining diverse feature types if they are handled by a model capable of integrating them equally. We also observe that the difference between validation score and testing score decreased by 0.4 compared to the Baseline RF model. This means we sucessfully reduced overfitting.

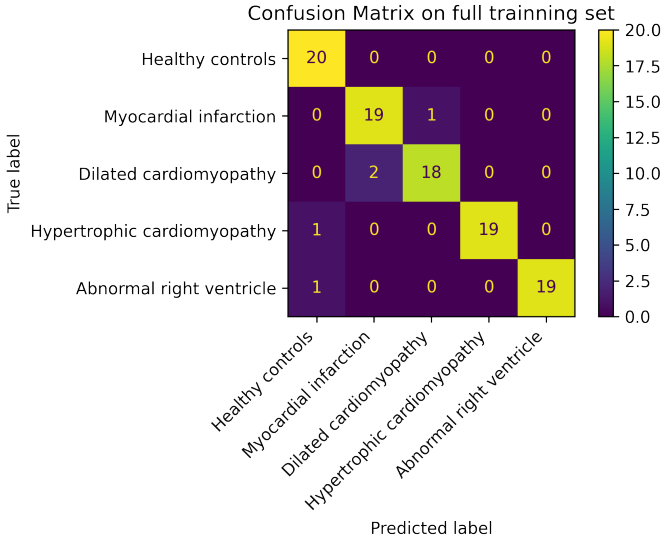


Figure 4.3: Confusion matrix on training set for the SVM model trained on 20 radiomics features.

Chapter 5

Conclusion and Perspectives

5.1 Conclusion and Perspectives

In this challenge, I explored multiple strategies for feature extraction, selection, and model fitting. Starting from classical shape descriptors, I progressively extended the feature space to include intensity and texture-based radiomics features. Along the way, I assessed the impact of univariate and multivariate selection techniques (ANOVA, mRMR), and compared different classifiers, including Random Forests and Support Vector Machines.

One key insight is that while shape features remain highly predictive, particularly due to their stability and their direct correlation with anatomical disease, they tend to dominate models like Random Forests, which select features greedily. In contrast, the SVM was more capable of leveraging complementary radiomics features, allowing intensity and texture information to contribute meaningfully to the final prediction.

Ultimately, the results obtained here show that integrating diverse radiomics descriptors,beyond shape, is feasible and beneficial, provided the model is adapted to handle them effectively. However, this process remains challenging. Despite incorporating GLCM and GLRLM texture features, I was not able to significantly improve classification accuracy compared to shape-only models. This may suggest that these descriptors are not optimal in their current form, or that they require a more complex model to express their full potential—such as deep learning architectures capable of capturing subtle local patterns.

I also analysed prediction probabilities from both Random Forests and SVMs, and noticed systematic confusion between Class 1 (myocardial infarction) and Class 2 (dilated cardiomyopathy). These two conditions are anatomically similar and likely hard to separate based on radiomics alone. When visualising samples in feature space, misclassified cases often appeared as “hard samples,” meaning they had fewer neighbours of their own class than of another. I attempted to mitigate this with SMOTE Borderline oversampling which oversample the "hard samples" , but without success.

Although somewhat disappointed that texture and signal-based features did not perform as well as expected, I remain convinced of their relevance. Their high dimensionality and noise sensitivity may simply require more sophisticated approaches, possibly involving feature learning from raw images. Future work could explore convolutional neural networks to extract higher-level representations from these domains.

Finally, another open question relates to feature redundancy: it is difficult to determine whether moderately correlated variables,often removed during selection,might actually enhance model robustness through ensembling effects. This reflection opened new perspectives for me, as I had not fully considered the trade-off between redundancy and complementarity before. I look forward to deepening my understanding of this aspect in future work.

Bibliography

[1] Irem Cetin et al. "A Radiomics Approach to Computer-Aided Diagnosis with Cardiac Cine-MRI". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2019.

[2] Loic Le Folgoc and Pietro Gori. *IMA205 Challenge 2025*. https://kaggle.com/competitions/ima-205-challenge-2025. Kaggle. 2025.

[3] Isabelle Guyon and André Elisseeff. "An Introduction to Variable and Feature Selection". In: *Journal of Machine Learning Research, Special Issue on Variable and Feature Selection* 3 (2003), pp. 1157–1182. doi: 10.1162/153244303322753616.

[4] Zahra Raisi-Estabragh et al. "Cardiac magnetic resonance radiomics: basic principles and clinical perspectives". In: *European Heart Journal - Cardiovascular Imaging* 21.4 (2020), pp. 349–356. doi: 10.1093/ehjci/jeaa028.

Appendix A

Heat Correlation Matrix

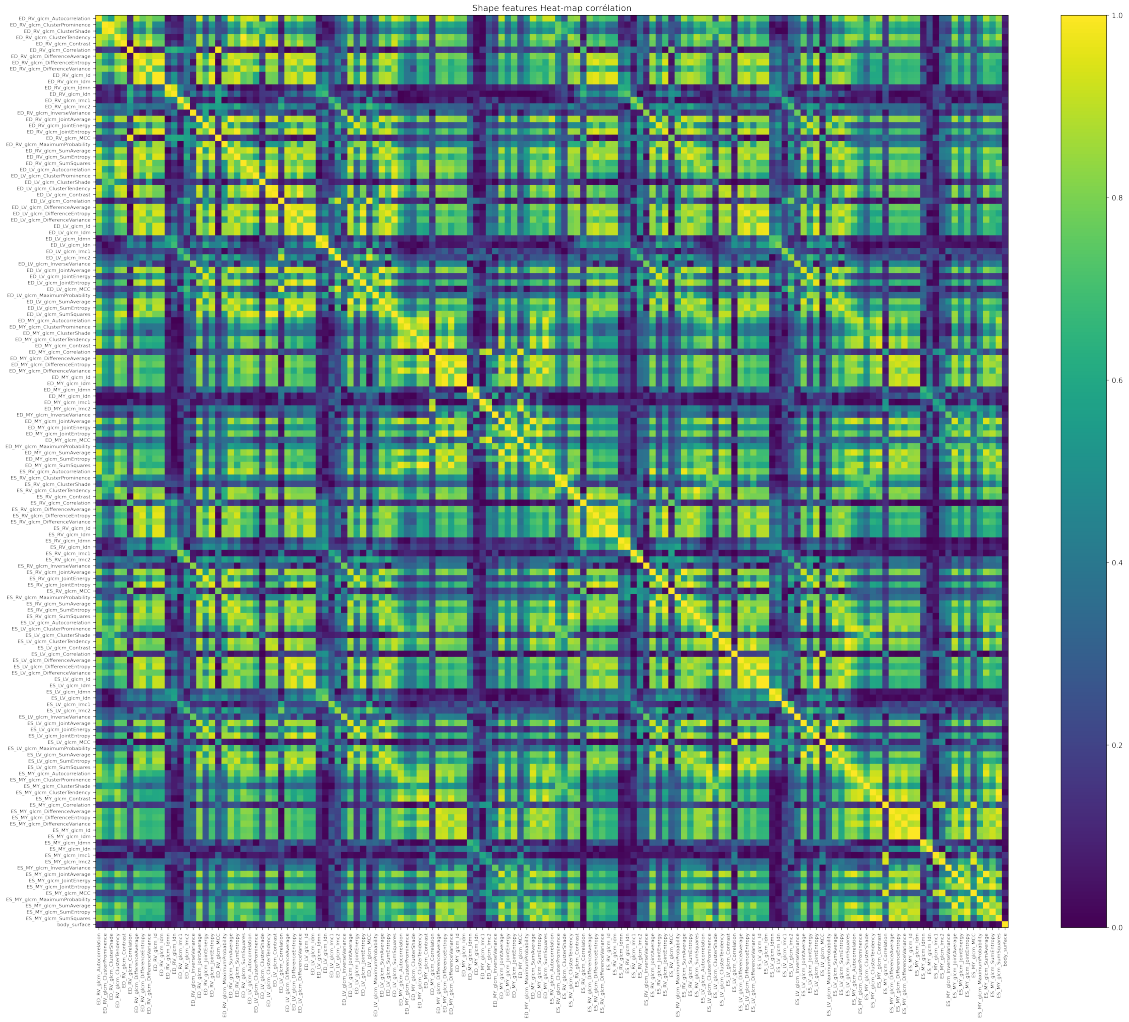


Figure A.1: GLCM features heat correlation map

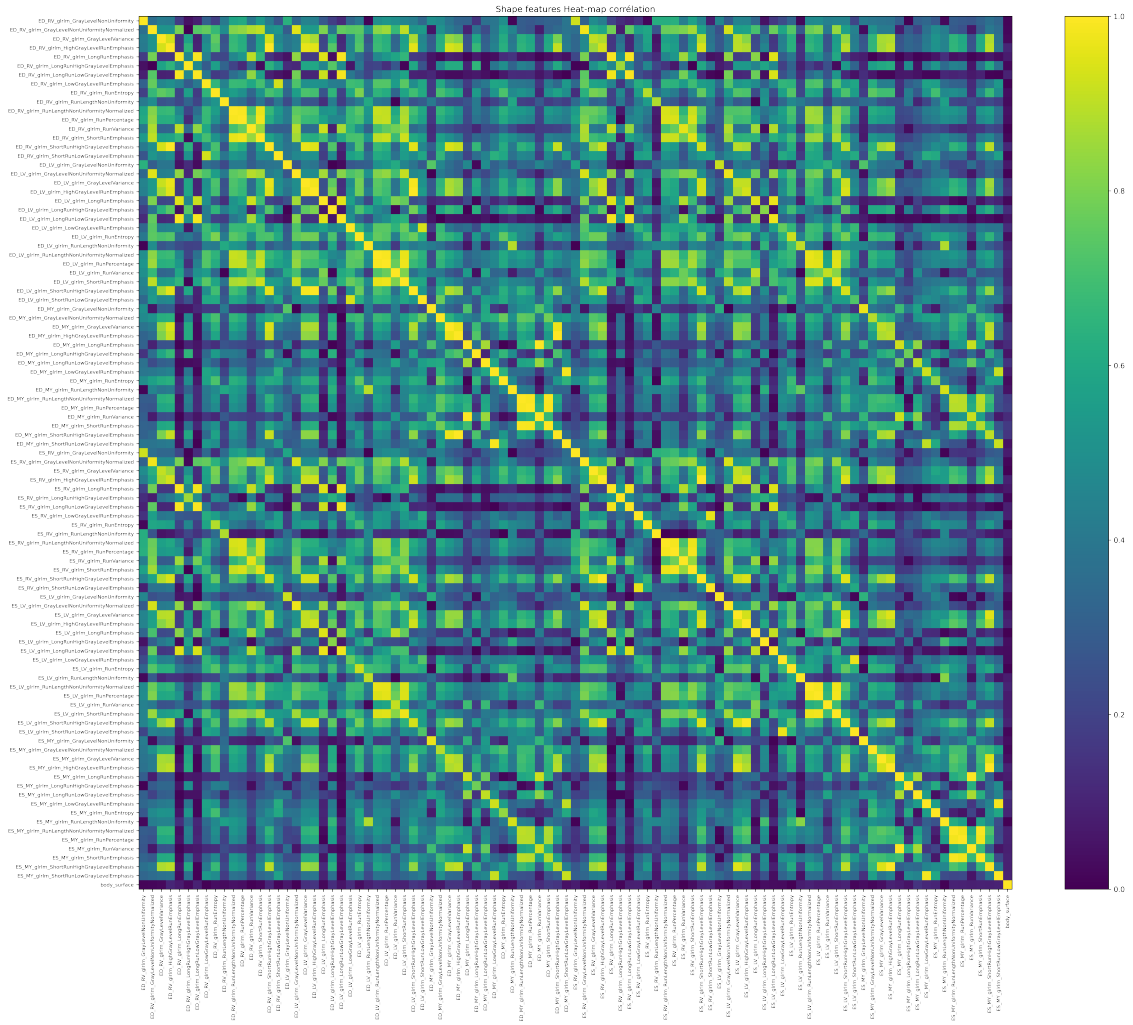


Figure A.2: GLRLM features heat correlation map