Investigating Cardiac Magnetic Resonance Imaging Feature Descriptors for Generalization of Cardiomyopathy Classification Models

Stephani S. H. Costa[1], Vagner Mendonça Gonçalves[1,2], Matheus A. O. Ribeiro[1], and Fátima L. S. Nunes[1]

[1]Laboratory of Computer Applications for Health Care,
Escola de Artes, Ciências e Humanidades, Universidade de São Paulo
Rua Arlindo Bettio, 1000, São Paulo – SP, Brazil, 03828-000

[2]Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, Campus São Paulo
Rua Pedro Vicente, 625, São Paulo – SP, Brazil, 01109-010

{stephani.henrique, vagner.goncalves, matheus.alberto.ribeiro, fatima.nunes}@usp.br

Abstract. *Supervised Machine Learning (SML) models can help physicians to compose more accurate diagnoses. Due to the diversity of machines, systems, and protocols, exams conducted at different centers may have high variability, decreasing the model generalization ability. This paper aims to evaluate generalization ability of SML models for cardiomyopathy classification in Cardiac Magnetic Resonance Imaging (CMRI), using a set of left ventricle morphological and motion features. We performed cross-validation tests on two public CMRI databases, comparing intra- and inter-dataset performances. The results are promising and demonstrate that the implemented features can contribute to building generalizable classification models.*

1. Introduction

Cardiovascular diseases are the leading cause of mortality in Brazil, accounting for approximately 1,100 deaths each day and it is estimated that more than 400 thousand Brazilians will die because of these diseases in 2025 [Sociedade Brasileira de Cardiologia 2025]. Among these diseases, cardiomyopathies, such as Dilated Cardiomyopathy (DCM) and Hypertrophic Cardiomyopathy (HCM), stand out due to their chronic and potentially fatal nature [Sundaram et al. 2021]. Early diagnosis is crucial for proper management and the expansion of treatment possibilities, directly impacting patient survival.

In recent years, classification based on Supervised Machine Learning (SML) models has been widely explored for the development and evaluation of Computer-Aided Diagnosis (CAD) approaches aiming at supporting the diagnosis of these diseases [Kagiyama, Tokodi, and Sengupta 2022]. However, due to the high variability of medical exams performed at different institutions caused by a diversity of machines, systems, and protocols, as well as the limitation of available datasets used in training and lack of descriptors that are insensitive to this variability, the generalization ability of a model may not be guaranteed for other datasets [Linardos et al. 2022].

Given this problem, this paper aims to evaluate generalization ability of SML models for cardiomyopathy classification in Cardiac Magnetic Resonance Imaging (CMRI), using a set of left ventricle morphological and motion features. The datasets were composed of features extracted from the LV segmentation in CMRI exams from two public databases. The main contribution presented in this paper is to demonstrate that LV morphological and motion features can contribute to the building of generalizable classification models.

2. Related Work

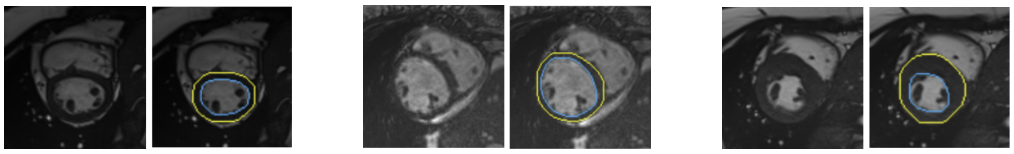
Diao et al. [2023] developed various models based on cardiac regions, segmented ventricles, and ventricular masks to distinguish cases of HCM and hypertensive heart disease. Zhou et al. [2023], in turn, employed SML models capable of distinguishing cases of DCM and ischemic cardiomyopathy. Zhang et al. [2023] proposed a radiomic model to automatically differentiate LV Non-compaction, HCM, and DCM, without requiring manual delineation. Strategies based on Transfer Learning and Federated Learning also have been explored. Linardos et al. [2022] investigated the use of Federated Learning for cardiomyopathy classification, while Sivaprasad et al. [2022] studied Transfer Learning models that can leverage anatomical features extracted from neural networks.

Although obtaining positive results, these approaches present significant challenges: federated methods can be costly due to the need to train multiple models on different devices; Deep Learning (DL) techniques require large volumes of labeled data and often result in less interpretable models; while transfer learning strategies may not generalize well to unseen domains. Unlike existing approaches, we propose the development of more robust features (regarding morphological and motion) reduced influence from variability between databases as a way to improve generalization ability.

3. Materials and Methods

We utilized two publicly available databases of CMRI exams: Automated Cardiac Diagnosis Challenge (ACDC) [Bernard et al. 2018] and Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation (M&Ms) [Martín-Isla et al. 2023]. From these databases, we selected cases of DCM, HCM, and No Anomaly (NA) diagnosis, corresponding to 90 (30 DCM, 30 HCM, 30 NA) and 244 (89 DCM, 75 HCM, 80 NA) cases of the ACDC and M&Ms databases, respectfully.

Each cine-CMRI exam consists of image sequences acquired at different time points throughout the cardiac cycle and across multiple slices of the heart. The main objects of interest in the analysis are the left ventricular (LV) cavity, defined as the area within the endocardium, and the myocardium, delineated between the epicardium and endocardium. Figure 1 illustrates these structures before and after the segmentation process.



(a) NA (b) DCM (c) HCM

Figure 1. Examples of cases for each diagnosis segmented by the automatic method used. The yellow and blue lines indicate the segmented outer borders of the myocardium and the heart chamber, respectively.

For the necessary script implementations, we used the Python programming language, version 3.11.5, as well as the open source libraries Scikit-Learn, version 1.3.0.

3.1. Classification Model Building Process

We applied the general classification model building process, structured into five subprocesses: A) Segmentation; B) Feature extraction; C) Feature normalization; D) Outer cross-validation (tests); E) Inner cross-validation (model and hyper-parameter tuning). Subprocess

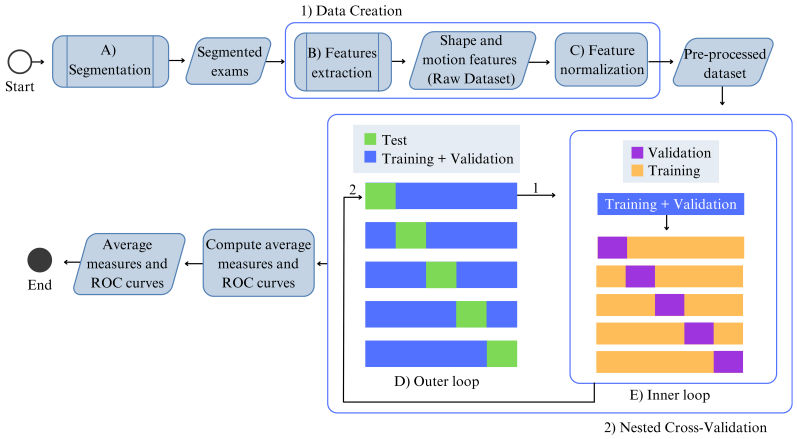


Figure 2. Classification model building process. The first step is the segmentation of the exams (Subprocess A), followed by the extraction of morphological and motion features (Subprocess B). Next, the features are normalized (Subprocess C) and used in a nested cross-validation(Subprocesses D and E).

D and E were conducted using a nested cross-validation strategy. Figure 2 illustrates the whole process, detailed next.

Segmentation (Subprocess A in Figure 2) : the segmentation was performed by an automatic segmentation pipeline composed of preprocessing and DL techniques. The preprocessing stage is composed by image histogram equalization and region of interest extraction, while the DL is trained to segment the LV cavity and myocardium regions. The full segmentation pipeline is presented by Ribeiro, Gutierrez, and Nunes [2023].

Feature Extraction and Normalization (Subprocesses B and C in Figure 2): We represented each CMRI exam through a vector of 60 features in total, comprising both morphological and motion features, accompanied by its target class (the diagnosis associated with the exam). These features were extracted from the set of segmented CMRI exams using feature descriptors that performs calculations based on pixel values. These calculations include geometric measurements and temporal variation analysis, allowing for the characterization of cardiac structures and their movements over time. All extracted features underwent Min-Max normalization to scale the features within the range [0.0, 1.0], reducing scale variability.

Nested Cross-validation (Subprocesses D and E in Figure 2) : In the Outer Cross-Validation (CV) the dataset was divided using a 5-Fold CV strategy. In each Outer CV round, one fold was set aside for testing, while the union of the remaining four folds was further divided into 5 folds for training and validation in the Inner loop. For each Outer CV round, the pipeline applied to evaluate the classifier involved steps for class balancing, dimensionality reduction (feature selection and transformation), and Classifier Induction Algorithm (CIA) training. The best combination of algorithms for each step, as well as the hyper-parameters values applied to train the CIA were selected by performing a Grid Search process based on the Inner CV and selection the one that maximizes the weighted macro-average F_1-score.

3.2. Experimental Setup

We designed a classification pipeline by evaluating different combinations of algorithms, dataset balancing strategies, feature selection methods, and data transformation techniques. The complete set of strategies applied is detailed in Table 1.

We evaluated the classification models' performance when trained and tested on the

Table 1. Algorithms and strategies applied in this research. LDA: Linear Discriminant Analysis; RF: Random Forest; SVC: Support Vector Classifier; XGB: Extreme Gradient Boosting; PCA: Principal Component Analysis; SMOTE: Synthetic Minority Over-sampling Technique.

CIA	Strategy of class balancing	Strategy of feature selection	Strategy of feature transformation
LDA RF SVC XGB	None Random Undersampling SMOTE	None SelectKBest Variance Threshold	None LDA PCA

ACDC and M&Ms datasets. Additionally, we assessed the generalization ability of the descriptors by comparing the results obtained when the model was trained and tested on the same dataset to those achieved when training on one dataset and testing on the other.

As main performance metrics, we used the weighted macro-averaged F_1-score, precision, recall, and accuracy, as well as Area Under the ROC Curve (AUC). For analyzing the multiclass confusion matrix, we applied the One-versus-the-Rest approach.

4. Results and Discussion

Tables 2 and 3 show the results of the best combinations of the pipeline mentioned in Section 3.2 for each algorithm, when trained on M&Ms and tested on ACDC (Table 2) and when trained on ACDC and tested on M&Ms (Table 3).

Table 2. Mean performance values and respective standard deviations achieved by the best classification model per CIA when trained on the M&Ms dataset. The best performance for each metric is highlighted in bold.

CIA	Testing Dataset	F_1-score	Precision	Recall	Accuracy	AUC
LDA	ACDC	0.64 ± 0.17	0.74 ± 0.19	0.69 ± 0.14	0.69 ± 0.14	0.90 ± 0.12
	M&Ms	0.72 ± 0.05	0.74 ± 0.04	0.72 ± 0.05	0.72 ± 0.05	0.88 ± 0.05
RF	ACDC	0.77 ± 0.03	0.81 ± 0.01	0.79 ± 0.02	0.79 ± 0.02	0.96 ± 0.01
	M&Ms	0.72 ± 0.1	0.73 ± 0.09	0.72 ± 0.1	0.72 ± 0.1	0.86 ± 0.06
XGB	ACDC	0.67 ± 0.18	0.73 ± 0.15	0.70 ± 0.14	0.70 ± 0.14	0.89 ± 0.1
	M&Ms	0.68 ± 0.08	0.68 ± 0.08	0.68 ± 0.09	0.68 ± 0.08	0.85 ± 0.04
SVC	ACDC	$\mathbf{0.85 \pm 0.03}$	$\mathbf{0.87 \pm 0.03}$	$\mathbf{0.85 \pm 0.03}$	$\mathbf{0.85 \pm 0.03}$	$\mathbf{0.97 \pm 0.01}$
	M&Ms	0.72 ± 0.07	0.73 ± 0.06	0.72 ± 0.07	0.72 ± 0.07	0.88 ± 0.05

Table 3. Mean performance values and respective standard deviations achieved by the best classification model per CIA when trained on the ACDC dataset. The best performance for each metric is highlighted in bold.

CIA	Testing dataset	F_1-score	Precision	Recall	Accuracy	AUC
LDA	M&Ms	0.51 ± 0.03	0.75 ± 0.01	0.55 ± 0.02	0.54 ± 0.02	0.84 ± 0.01
	ACDC	$\mathbf{0.94 \pm 0.05}$	$\mathbf{0.96 \pm 0.04}$	$\mathbf{0.94 \pm 0.05}$	$\mathbf{0.94 \pm 0.05}$	$\mathbf{0.99 \pm 0.01}$
RF	M&Ms	0.52 ± 0.16	0.64 ± 0.18	0.56 ± 0.12	0.56 ± 0.1	0.74 ± 0.11
	ACDC	0.90 ± 0.07	0.91 ± 0.06	0.90 ± 0.06	0.90 ± 0.06	0.96 ± 0.04
XGB	M&Ms	0.50 ± 0.10	0.68 ± 0.05	0.54 ± 0.07	0.53 ± 0.08	0.76 ± 0.04
	ACDC	0.88 ± 0.10	0.89 ± 0.09	0.88 ± 0.10	0.88 ± 0.10	0.95 ± 0.04
SVC	M&Ms	0.57 ± 0.03	0.74 ± 0.01	0.58 ± 0.02	0.58 ± 0.02	0.86 ± 0.02
	ACDC	0.93 ± 0.05	0.94 ± 0.05	0.93 ± 0.05	0.93 ± 0.05	0.98 ± 0.03

The results indicate that the performance of the algorithms varies considerably depending on the training and test datasets. As expected, models trained on the larger dataset (M&Ms) and tested on the smaller one (ACDC) performed better, with most models achieving an accuracy above 0.70. In the reverse scenario, where the models were trained on ACDC and tested on M&Ms, performance was lower, with all models recording an accuracy below 0.60.

Overall, for the models trained on the M&Ms dataset, the SVC algorithm performed the best, achieving the highest values in all evaluation metrics. When trained on M&Ms and

tested on ACDC, this model achieved an average accuracy of 0.85, 0.87 precision, as well as superior F_1-score and recall compared to the other methods evaluated. RF also showed robust performance in this configuration, with an average recall and accuracy of 0.79, along with good precision and F_1-score values.

On the other hand, when the models were trained on ACDC and tested on M&Ms, there was a significant drop in performance. SVC achieved an accuracy of 0.58, while RF achieved 0.56, both showing similar reductions in other metrics. This decline occurs because the ACDC dataset has less diversity compared to M&Ms, which limits the models' ability to generalize to new examples. LDA followed a similar trend, achieving an accuracy of 0.69 when trained on M&Ms and tested on ACDC, but only 0.54 in the reverse configuration.

Since M&Ms has a larger amount of data and greater variability in the images, models trained on it can learn more general patterns, enabling more robust performance when tested on ACDC. On the other hand, models trained on ACDC, being a smaller and less diverse dataset, fail to capture the full complexity present in M&Ms, resulting in inferior performance when applied to this test set.

Despite the drop in performance measures when the training and test sets were swapped, the algorithms still maintained a considerable level of performance. This suggests that the descriptors used have discriminative potential, provided they are trained on a sufficiently diverse dataset. The results reinforce the importance of using comprehensive training datasets to improve model generalization in real-world applications.

The achieved performance measures are in line with related studies. Izquierdo et al. [2021] achieved 0.95 AUC using only M&Ms, while our SVC reached 0.97 in a cross-dataset setup. Zhang et al. [2023] trained with ACDC and M&Ms combined, reaching 0.912 accuracy, while our cross-dataset test resulted in 0.85. Atehortúa, Romero, and Garreau [2022] achieved 0.86 precision, similar to our model's 0.87. However, it is important to highlight that differences in datasets and protocols make direct comparisons difficult.

Analyzing the top 40 models, it was observed that none applied data balancing, and the most common strategies included the use of the ranking algorithm based on mutual information for feature selection (75%) and PCA as a transformation technique (37.5%). More than 80% of the most frequent features were morphological — primarily ventricular volumes — while features related to endocardial motion had less impact on classification.

Although this study can be expanded, the results obtained suggest that it is possible to achieve generalizable cardiomyopathy classification models in CMRI exams with the application of morphological and motion features with low sensitivity to variability between databases, without the need for other unnecessarily complex approaches.

5. Conclusion

Our results indicate that models trained on M&Ms database generalize better to ACDC database than the other way around, reinforcing the importance of a larger and more diverse training set. SVC proved to be the most robust algorithm, consistently achieving the best performance across all evaluation metrics. We also show that, with sufficiently diverse data, the morphological and motion descriptors exhibit strong discriminative ability on their own, allowing models to learn relevant patterns even with different test datasets.

As future work, we intend to include new databases in the evaluation of model generalization, as well as apply specific classification model generalization techniques to assess

their impact on the proposed approach. Additionally, we plan to identify a subset of features that consistently enhances classification performance.

Acknowledgements

This work was supported in part by the São Paulo Research Foundation (FAPESP) through a undergraduate research scholarship [grant number 2024/13568-0], the National Institute of Science and Technology – Medicine Assisted by Scientific Computing (INCT-MACC) [grant number 2014/50889-7], the National Council for Scientific and Technological Development (CNPq) [grant numbers 307710/2022-0], and the Office of the USP's Vice-Provost for Research and Innovation.

References

Atehortúa, A., Romero, E., and Garreau, M. (2022). Characterization of motion patterns by a spatio-temporal saliency descriptor in cardiac cine MRI. *Computer Methods and Programs in Biomedicine*, 218(106714).

Bernard, O. 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.

Diao, K. et al. (2023). Multi-channel deep learning model-based myocardial spatial–temporal morphology feature on cardiac MRI cine images diagnoses the cause of LVH. *Insights into Imaging*, 14(70).

Izquierdo, C. et al. (2021). Radiomics-based classification of left ventricular non-compaction, hypertrophic cardiomyopathy, and dilated cardiomyopathy in cardiovascular magnetic resonance. *Frontiers in Cardiovascular Medicine*, 8(764312).

Kagiyama, N., Tokodi, M., and Sengupta, P. P. (2022). Machine learning in cardiovascular imaging. *Heart Failure Clinics*, 18(2):245–258.

Linardos, A. et al. (2022). Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease. *Scientific Reports*, 12(3551).

Martín-Isla, C. et al. (2023). Deep learning segmentation of the right ventricle in cardiac MRI: the M&Ms challenge. *IEEE Journal of Biomedical and Health Informatics*, 27(7):3302–3313.

Ribeiro, M. A. O., Gutierrez, M. A., and Nunes, F. L. S. (2023). Improving deep learning shape consistency with a new loss function for left ventricle segmentation in cardiac MRI. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems*.

Sivaprasad, R. et al. (2022). Heart disease prediction and classification using machine learning and transfer learning model. In *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pages 595–601.

Sociedade Brasileira de Cardiologia (2025). Cardiômetro: monitoramento de mortes por doenças cardiovasculares no Brasil. Available at: `http://www.cardiometro.com.br/`. Accessed: 4 Mar 2025.

Sundaram, D. S. B. et al. (2021). Natural language processing based machine learning model using cardiac MRI reports to identify hypertrophic cardiomyopathy patients. In *2021 Design of Medical Devices Conference*. The American Society of Mechanical Engineers.

Zhang, X. et al. (2023). Cardiac magnetic resonance radiomics for disease classification. *European Radiology*, 33(4):2312–2323.

Zhou, M. et al. (2023). Echocardiography-based machine learning algorithm for distinguishing ischemic cardiomyopathy from dilated cardiomyopathy. *BMC Cardiovascular Disorders*, 23(476).