Generalization of Cardiomyopathy Classification Models Based on Feature Descriptors from Magnetic Resonance Imaging

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 generalization ability of these models. 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 30% of annual deaths, according to the Ministry of Health [Ministério da Saúde 2023]. According to the Brazilian Society of Cardiology, more than 1,100 people die each day from heart diseases, 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 [Xia et al. 2019; 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 [Shen, Wu, and Suk 2017; Kagiyama, Tokodi, and Sengupta 2022]. However, there are still challenges related to the generalization of these models, especially when the databases used for training are from a single center, which can limit their applicability in different clinical contexts and practical applications [Ng et al. 2007; Linardos et al. 2022].

Due to the diversity of machines, systems, and protocols, medical exams performed at different institutions often exhibit variability regarding image resolution, contrast between different regions, presence of artifacts, and segmentation performed by specialists [Knoll et al. 2019]. Since data tend to remain isolated within institutions, many studies are limited to single-center databases for the development and evaluation of computational models applicable in CAD systems. As a consequence, the performance of a model may not be guaranteed for samples containing feature values significantly different from those present in the training set [Ng et al. 2007; Linardos et al. 2022]. Addressing this limitation requires strategies such as multi-center data training, domain adaptation, and robust feature selection methods [Zhang et al. 2023].

In this work, we propose to evaluate generalization ability of SML models for cardiomyopathy classification in CMRI, using a set of left ventricle morphological and motion features. The datasets were composed of features extracted from the LV segmentation of CMRI exams from two public databases. The main contribution presented in this paper is to demonstrate that morphological and motion feature descriptors can adequately represent the LV structure in CMRI exams from different databases in a generalizable way. This evidence can contribute to the construction of better generalizable classification models considering different sources.

The remainder of this paper is structured as follows: Section 2 presents related work; Section 3 describes materials and methods applied in the study; Section 4 presents and discusses the results; and Section 5 provides final considerations and future directions.

2. Related Work

The generalization of cardiomyopathy classification models from CMRI is still a challenge in current research. Some studies have proposed approaches to enhance the ability of SML and deep learning models to adapt to different domains and databases, reducing bias caused by variations in image acquisition and the studied population.

Diao et al. [2023] and Zhou et al. [2023], for example, conducted validations on datasets different from those used for model training and compared the effectiveness of different classification models based on discriminatory performance. 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 algorithms such as Random Forest (RF), Logistic Regression, Neural Networks, and XGBoost to generate SML models capable of distinguishing cases of DCM and ischemic cardiomyopathy. Both studies achieved positive results when testing their models on datasets from institutions different from those used for training.

Izquierdo et al. [2021] and Zhang et al. [2023] explored radiomic features extracted from CMRI images. The former study applied feature selection methods and classifiers to distinguish between HCM, DCM, and cases without characteristic anomalies of these diseases (NA, from No Anomalies), while the latter 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] and Goto et al. [2023] investigated the use of Federated Learning for cardiomyopathy classification, allowing models to be trained across different centers without direct data sharing, thus reducing the impact of institutional differences and promoting more robust generalization. Transfer Learning models, such as those studied by Sivaprasad et al. [2022], also showed improvements in adaptation to new datasets by leveraging anatomical features extracted from pre-trained neural networks.

Finally, the exploration of more robust architectures is a recurring theme. Wibowo et al. [2022] employed Convolutional Neural Networks and hybrid models to capture more complex patterns in CMRI images, improving generalization by incorporating multiple types of image-derived features. Atehortúa, Romero, and Garreau [2022], in turn, demonstrated that incorporating kinematic and morphological features improves model generalization, particularly when combining information from T1 mapping sequences with LV motion analysis.

Recent advances in cardiomyopathy classification have focused on reducing generalization bias through Federated Learning techniques, knowledge transfer, latent representations, data augmentation, and the development of more robust architectures. However, these approaches present significant challenges: federated methods can be costly due to the need to train multiple models on different devices; deep learning techniques require large volumes of labeled data and often result in less interpretable models; while knowledge transfer strategies may not generalize well to unseen domains. Although the generalization of classification models has received increasing attention in recent research on CAD systems for cardiomyopathies, the number of studies in this area remains relatively small. Unlike existing approaches, we propose the use of more domain-invariant features regarding morphological and motion as a way to obtain generalization. To achieve this, we developed feature descriptors that are minimally influenced by the variability of quality, resolution, and image collection parameters, allowing the model to better adapt to cases from different sources. This approach seeks to reduce dependence on external strategies and make models more applicable to diverse clinical scenarios.

3. Materials and Methods

This section presents the materials and methods applied in this study. Section 3.1 describes the public databases of CMRI exams used for training and testing the classification models. Section 3.2 details the process of building classification models. Finally, Section 3.3 presents the experimental setup defined for the study.

3.1. Materials

To evaluate the generalization ability of SML models in the classification of cardiomyopathies, 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 corresponding to DCM, HCM, and NA subjects. The ACDC database includes a total of 150 real clinical exams acquired at the University Hospital of Dijon (France), of which 90 correspond to the pathologies explored in this study, distributed as 30 DCM, 30 HCM, and 30 NOR cases. The M&Ms database comprises 345 cases from different machines and centers, of which 244 serve the research interests, including 89 DCM, 75 HCM, and 80 NOR cases. In total, 334 cine-CMRI exams were obtained from multiple clinical centers and different scanner vendors, ensuring diversity in imaging conditions.

Each cine-CMRI exam consists of image sequences acquired at different time points throughout the cardiac cycle and across multiple slices of the heart. These images capture cardiac dynamics, covering the stages of diastole (relaxation and blood filling) and systole (contraction and blood pumping). 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.These structures play a crucial role in assessing different cardiac conditions and are fundamental for the segmentation and classification of cardiomyopathies.

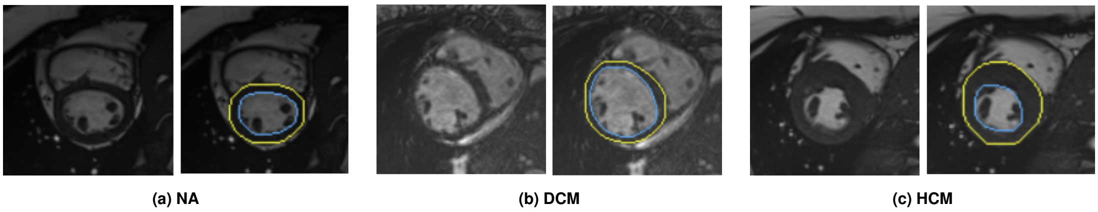

Figure 1. Examples of cases for each diagnosis segmented by the automatic method used. The yellow and blue lines indicate the segmented outer borders of the myocardium and the heart chamber, respectively.

In the next section, we detail the process of building the classification model. For the necessary script implementations, we used the Python programming language, version 3.11.5, as well as the open source libraries Scikit-Learn [Pedregosa et al. 2011], version 1.3.0, and Imbalanced-Learn [Lemaître, Nogueira, and Aridas 2017], version 0.11.0.

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

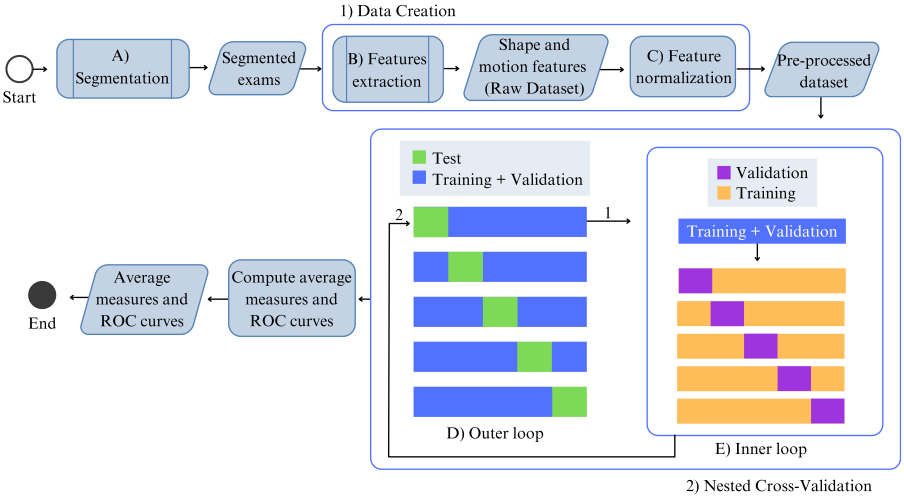

Figure 2. Classification model building process. The first step is the segmentation of the exams (Subprocess A), followed by the extraction of morphological and motion features (Subprocess B). Next, the features are normalized (Subprocess C) and used in a nested cross-validation(Subprocesses D and E).

3.2.1. Segmentation

The segmentation of LV structures (Subprocess A in Figure 2) was performed by an automatic segmentation pipeline composed of preprocessing, deep learning techniques, and postprocessing stages. The preprocessing stage is composed by image histogram equalization and region of interest extraction, useful for reducing image-related variability [Ribeiro and Nunes 2021]

In the segmentation stage, an U-net network is trained to segment the LV cavity and myocardium regions, while in the postprocessing stage, common anatomical errors, such as the presence of holes and multiple regions, are fixed. The full segmentation pipeline is presented in Ribeiro, Gutierrez, and Nunes [2023]. Examples of produced segmentations for each diagnosis is presented in Figure 3. The same segmentation pipeline was used to segment all used exams, ensuring consistency in the extracted features across different imaging sources.

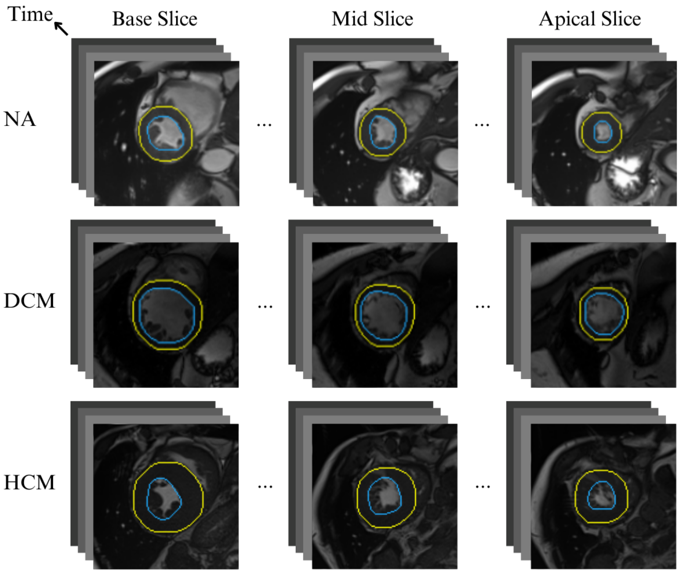

Figure 3. Examples of CMRI exams for each diagnosis segmented by the automatic method used. The yellow and blue lines indicate, respectively, the segmented LV epicardial and endocardial contours obtained from the myocardium and LV cavity segmentation.

3.2.2. Feature Extraction and Normalization

We represented each CMRI exam through a vector of 60 features in total, comprising both morphological and motion characteristics, 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 (Subprocess B in Figure 2). These calculations include geometric measurements and temporal variation analysis, allowing for the characterization of LV structure and movement over time. These newly developed features aim to refine the analysis of cardiac dynamics, providing a more generalizable representation of structural and functional changes across different cardiac pathologies. A description of each feature can be found in Table 1.

The extracted features underwent Min-Max normalization (Subprocess C in Figure 2) to scale the features within the range [0.0, 1.0], ensuring that all values remained within a comparable magnitude and improving model performance.

3.2.3. Nested Cross-validation

In the Outer Cross-Validation (CV) (Subprocess D in Figure 2), 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 used for training (outer training fold).

Table 1. Features computed using the new descriptors developed in this research.

Features	ID	Description
Endocardial Motion Analysis	0 to 2	Endocardial motion across frames for basal, mid, and apical slices during systole.
	11 to 13	Endocardial motion across frames for basal, mid, and apical slices during diastole.
	22 to 24	Endocardial motion across frames for basal, mid, and apical slices throughout the entire cardiac cycle.
Mean and Standard Deviation of Endocardial Motion	3 to 4	Mean and standard deviation of endocardial motion during systole, evaluated over the entire interval, at mid-systole, and at end-systole.
	14 to 15	Mean and standard deviation of endocardial motion during diastole, evaluated over the entire interval, at mid-diastole, and at end-diastole.
	25 to 26	Mean and standard deviation of endocardial motion over the entire cardiac cycle.
Myocardial and LV Cavity Areas	5 to 8	Mean myocardial area per slice at mid-systole and end-systole.
	9 to 10	Mean LV cavity area at end-systole (maximum contraction).
	16 to 19	Mean myocardial area per slice at mid-diastole and end-diastole.
	20 to 21	Mean LV cavity area at end-diastole (maximum relaxation).
	27 to 28	Difference in mean and standard deviation of LV cavity area between end-diastole (maximum relaxation) and end-systole (maximum contraction).
LV Cavity and Myocardial Volumes	29 to 32	LV cavity volume and myocardial volume at mid-systole and end-systole.
	35 to 38	LV cavity volume and myocardial volume at mid-diastole and end-diastole.
	33 to 34	Difference between maximum and minimum LV cavity volume and myocardial volume during systole.
	39 to 40	Difference between maximum and minimum LV cavity volume and myocardial volume during diastole.
	41 to 43	Difference between maximum and minimum LV cavity volume and myocardial volume throughout the entire cardiac cycle, including ejection fraction calculation.
Perimeter of the Endocardium and Myocardium	44 to 51	Mean and standard deviation of endocardial and myocardial perimeters at mid-systole and end-systole, for basal, mid, and apical slices.
	52 to 59	Mean and standard deviation of endocardial and myocardial perimeters at mid-diastole and end-diastole, for basal, mid, and apical slices.

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 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 (Subprocess E in Figure 2), using the respective outer training fold. Grid Search is a brute-force approach to identify, within a set of predefined options, the algorithms, hyper-parameters and strategies that maximize a target performance measure [Bergstra and Bengio 2012]. More details on the techniques and algorithms used are provided in Section 3.3

For the Inner CV subprocess, a new 5-fold CV strategy was applied on the outer training fold. One Inner CV was performed for each combination among algorithms/strategies for class balancing, feature selection, and feature transformation, as well as CIA hyper-parameter values. The Grid Search was configured to search the combination that maximize the weighted macro-average F_1-score.

Once the final pipeline was determined by the Grid Search process, the classification model of the corresponding Outer CV round was retrained on the outer training fold and tested on the outer test fold. The final classification model of the Outer CV round was also tested on an external dataset to assess its generalization ability.

3.3. Experimental Setup

We defined several algorithms, strategies, and hyper-parameter values to tune the pipeline for building classification models. Four CIA were evaluated: Linear Discriminant Analysis (LDA) [Hastie, Tibshirani, and Friedman 2009]; RF [Breiman 2001]; Support Vector Classifier (SVC) [Cortes and Vapnik 1995]; and Extreme Gradient Boosting (XGB) [Chen and Guestrin 2016].

To handle dataset imbalance, we applied one of three distinct strategies for class balancing: no modification (None); Synthetic Minority Over-sampling Technique (SMOTE) [Chawla et al. 2002]; or Random Undersampling.

Three feature selection approaches were considered: no selection (None); SelectKBest [Pedregosa et al. 2011], which selects the most relevant features based on statistical tests using mutual information [Kraskov, Stögbauer, and Grassberger 2004]; and, Variance Threshold [Gheyas and Smith 2010], which removes features with low variability.

We also explored data transformation techniques. Three approaches were tested: no transformation (None); LDA as a supervised dimensionality reduction method [Hastie, Tibshirani, and Friedman 2009]; and, Principal Component Analysis (PCA) as an unsupervised method [Jolliffe 2002]. The impact of these transformations was analyzed in conjunction with the other strategies, exploring all the combinations presented in the Table 2.

A total of 294 combinations were tested across balancing strategies, feature selection, and transformation. Considering the possible hyperparameter combinations tested for each algorithm, in each cross-validation round, the CIA with the fewest tested combinations was evaluated through 14,114 configurations, while the most tested CIA was evaluated with 172,801 through combinations.

Table 2. Algorithms and strategies applied in this research.

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

We evaluated the classification models' performance when trained and tested on the 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. Our descriptors generate measurements based on the resolution information available in the exam metadata. These measurements are computed in the same unit across all datasets, which allows for better generalization ability of the models.

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 confusion matrix in the context of the multiclass problem, we applied the One-versus-the-Rest approach. The metrics used are traditional in classification studies and were chosen based on the literature [Moreno, Rodriguez, and Martínez 2019; Izquierdo et al. 2021; Liu et al. 2023].

4. Results and Discussion

This section presents the results of the classification experiments. Section 4.1 describes the overall performance of the models trained and tested on the ACDC and M&Ms datasets, analyzing the generalization ability of the features by comparing intra-dataset and cross-dataset evaluation scenarios. Then, Section 4.2 discusses the influence of data balancing and dimensionality reduction on the results.

4.1. General performance

Tables 3 and 4 show the results of the best combinations of the pipeline mentioned in Section 3.3 for each algorithm, when trained on M&Ms and tested on ACDC (Table 3) and when trained on ACDC and tested on M&Ms (Table 4).

Table 3. Mean performance and respective standard deviation computed from the best classification model obtained in each holdout per algorithm and dataset 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 4. Mean performance and respective standard deviation computed from the best classification model obtained in each holdout per algorithm and dataset 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. These results suggest that the quantity and diversity of training data have a significant impact on the model's generalization ability.

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 F1-score and recall compared to the other methods evaluated. More specifically, SVC achieved a recall of 0.85, a good hit rate for pathological cases. RF also showed robust performance in this configuration, with an average recall and accuracy of 0.79, along with good precision and F1-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. This justifies such high results in the metrics when classification models are trained and tested on ACDC, which, being a smaller dataset, allows the network to fit the data very well but does not develop an adequate generalization ability.

The AUC analysis confirms the trends observed in the other metrics. SVC stood out as the best model, achieving an average AUC of 0.97 when trained on M&Ms and tested on ACDC, indicating the model's discriminative capacity. RF also performed strongly, with an AUC of 0.96. However, when trained on ACDC and tested on M&Ms, both models showed a considerable reduction, with SVC reaching 0.86 and RF 0.74.

As shown, AUC measures are numerically superior to F1-score, precision, and recall, as they are computed based on the default operating point of classification algorithms in Scikit-Learn [Manning, Raghavan, and Schütze 2008]. This suggests that an appropriate selection of the operating point can significantly enhance classifier performance.

For the models trained on the ACDC dataset, a significant performance drop was observed when comparing the results of testing the model on the test fold of the training set itself with the results of testing the model on the external dataset. While models trained and tested on ACDC showed an accuracy above 0.90, the results of testing the same models on an external dataset did not even reach an average accuracy of 0.60 or an F1-Score above 0.60.

Despite the drop in AUC and other metrics 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 highest mean performance measures achieved by the classification models tested in this study are in line with results from 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] evaluated cross-dataset generalization and reported 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.

4.2. Influence of balancing and dimensionality reduction techniques

Evaluating the strategies employed in the top 40 models, it was observed that data balancing was not applied in any case, indicating that the models performed well even without the use of specific techniques to address potential imbalances. Regarding feature selection, 75% of the models preferred to use the ranking algorithm based on mutual information. A further analysis of the feature selection choices revealed that more than half of the models (52.8%) preferred to select 30 or fewer features, corresponding to 50% of the total available features. As for transformations, half of the models (50%) did not employ any, but among those that did, PCA was used in 37.5% of the models.

The feature selection analysis revealed a predominance of morphological attributes, such as volumes and areas of the left ventricle, with more than 80% of the 20 most frequent features belonging to this category. These represent cavity and myocardial volumes throughout the cardiac cycle and the variation between phases. These results indicate that morphological descriptors play a central role in classification, being consistently selected in the best-performing models.

Most models demonstrated good generalization ability by selecting between 20 and 30 features from the original dataset. This suggests that, although the features we explore provide good performance, some of them can be redundant, making their use unnecessary for cardiomyopathy classification.

Another relevant aspect was the difference in behavior between the datasets analyzed. Models trained with the M&Ms dataset showed greater consistency in choosing the feature selection and transformation techniques that improved performance. In contrast, models trained with the smaller ACDC dataset exhibited higher variability for these pipeline steps, with different strategies maximizing results in each round of cross-validation. This may indicate that when a larger number of training samples are available, the tendency is to achieve a better sample representation of the problem universe. Consequently, it is reasonable for the tuning process to select the same techniques and similar hyperparameter values in the different tests performed for the same CIA. On the other hand, when using small training sets, the sample representation of the problem's universe is limited, which may explain the greater variability in the techniques and hyperparameter values selected by the tuning process across different tests.

5. Conclusion

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 that are minimally influenced by the variability of quality, resolution, and image collection parameters, without the need for other unnecessarily complex approaches.

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. RF also stood out, especially when trained on M&Ms.

As a limitation of our study, we highlight that our initial study only includes two distinct datasets for working on generalization, one of which is small and not very diverse. As future work, we plan to include new databases in the evaluation of model generalization and apply specific generalization techniques to assess their impact on the proposed approach. Additionally, we intend to further investigate the impact of variations in the techniques used at different stages of the pipeline, as well as 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):1–12.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305.

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.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, 16(1):321–357.

Chen, T. and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

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):1–11.

Gheyas, I. A. and Smith, L. S. (2010). Feature subset selection in large dimensionality domains. *Pattern Recognition*, 43(1):5–13.

Goto, S. et al. (2023). Multinational federated learning approach to train ECG and echocardiogram models for hypertrophic cardiomyopathy detection. *Circulation*, 146(10):755–769.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer, 2nd edition.

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):1–10.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer, 2nd edition.

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

Knoll, F. et al. (2019). Assessment of the generalization of learned image reconstruction and the potential for transfer learning. *Magnetic Resonance in Medicine*, 81(1):116–128.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69(6):066138–1–066138–16.

Lemaître, G., Nogueira, F., and Aridas, C. K. (2017). Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

Linardos, A. et al. (2022). Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease. *Scientific Reports*, 12(3551):1–12.

Liu, Q. et al. (2023). Papillary-muscle-derived radiomic features for hypertrophic cardiomyopathy versus hypertensive heart disease classification. *Diagnostics*, 13(9):1544:1–1544:15.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England.

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.

Ministério da Saúde (2023). “Use o coração para vencer as doenças cardiovasculares”: 29/9 – Dia Mundial do Coração. Available at: `https://bvsms.saude.gov.br/use-o-coracao-para-vencer-as-doencas-cardiovasculares-29-9-dia-mundial-do-coracao/`. Accessed: 4 Mar 2025.

Moreno, A., Rodriguez, J., and Martínez, F. (2019). Regional multiscale motion representation for cardiac disease prediction. In *2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA)*, pages 1–5. IEEE.

Ng, W. W. Y. et al. (2007). Image classification with the use of radial basis function neural networks and the minimization of the localized generalization error. *Pattern Recognition*, 40(1):19–32.

Pedregosa, F. et al. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830.

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 (CBMS)*. IEEE.

Ribeiro, M. A. O. and Nunes, F. L. S. (2021). Evaluating the pre-processing impact on the generalization of deep learning networks for left ventricle segmentation. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3505–3512. IEEE.

Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248.

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. IEEE.

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*, pages V001T03A005–1–V001T03A005–5. The American Society of Mechanical Engineers.

Wibowo, A. et al. (2022). Cardiac disease classification using two-dimensional thickness and few-shot learning based on magnetic resonance imaging image segmentation. *Journal of Imaging*, 8(7):194:1–194:16.

Xia, C. et al. (2019). A multi-modality network for cardiomyopathy death risk prediction with CMR images and clinical information. In Shen, D. et al., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, volume 11765 of *Lecture Notes in Computer Science*, pages 577–585. Springer Nature.

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):1–10.