DELIVERABLE

Project Acronym: SMASH-HCM

Grant Agreement number: 101137115

Project Title: Stratification, Management, and Guidance of Hypertrophic Cardiomyopathy Patients using Hybrid Digital Twin Solutions

D6.4 – Initial Data-Driven Model.

Revision: 1.0

Authors and Contributors	Name (beneficiary); POLIMI			
Responsible Author	Mainardi Luca		Email	luca.mainardi@polimi.it
	Beneficiary	POLIMI	Phone	+39-335-7996974

Project co-funded by the European Commission within HORIZON-HLTH-2023-TOOL-05-03		
Dissemination Level		
PU	Public, fully open	X
CO	Confidential, restricted under conditions set out in Model Grant Agreement	
CI	Classified, information as referred to in Commission Decision 2001/844/EC	

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Funded by the European Union

Revision History, Status, Abstract, Keywords, Statement of Originality

Revision History

Revision	Date	Author	Organisation	Description
0.1	4/7/25	LM	POLIMI	Skeleton
0.2	10/7/25	MT	POLIMI	ML model included
0.7	10/8/25	FA, MT, CGC, MvG, KC	POLIMI, UCD, PHARM, TUNI	Advanced draft
0.9	1/9/25	LM	POLIMI	First draft
1.0	1/9/25	ALL	ALL	Final version

Date of delivery	Contractual:	30.09.2025	Actual:	30.09.2025
Status	final ☒ /draft ☐			

Abstract (for dissemination)	This Deliverable presents the predictive data-driven models developed within Task T6.3, aimed at enriching biomarkers for hypertrophic cardiomyopathy (HCM) progression and comorbidities, and at enabling clustering and prediction methods for patient stratification, disease progression, and treatment response. The models are built on ECG, CMR, echocardiography, and multi-organ data, and are designed to complement those reported in Deliverable D6.2. Their integration with mechanistic models (planned in Task T6.4) is expected to enhance predictive power and clinical relevance. The document outlines the available datasets and the adopted data harmonization strategy, describes the developed models from single-modality to multi-modality approaches, and concludes with an overview of progress and next steps.
Keywords	Data-driven models, HCM patient stratification, multi-modality approaches

Statement of originality

This Deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of Content

Executive Summary

This Deliverable describes the data-driven predictive models developed within Task T6.3, with the objectives of: a) enriching the pool of predictive biomarkers for HCM progression and comorbidities using electrocardiogram (ECG) signals, cardiac magnetic resonance (CMR) and echocardiography images, and multi-organ data; and b) developing multilevel clustering and prediction methods for patient stratification, disease progression forecasting, and treatment response assessment.

The models presented here are intended to complement the library of models described in Deliverable D6.2 (obtained through literature mining), which will be implemented as part of Task T6.2. Their integration, planned in Task T6.4, with mechanistic models (WPs 3–5) is expected to enhance patient stratification and improve treatment response prediction in HCM.

Since the development of new models relies on the availability of data, two datasets have so far been provided within the project and have formed the basis for model development. However, as only partial data have been released to date, this Deliverable presents preliminary results based on the data available at the time of writing.

The document is organized as follows:

- Section 1 provides a brief overview of the datasets made available by the UNIFI and RENNES partners. It also outlines the strategy adopted to harmonize data usage for training and validation, to minimize bias and enable meaningful comparison across models. This required establishing a clear separation between training, validation, and test datasets.
- Section 2 describes the developed models and presents preliminary results obtained on the datasets introduced in Section 1. The models are presented according to the complexity of their inputs: starting from single-modality approaches (ECG, CMR, or echocardiography-based models) and progressing toward multi-modality models that integrate information across modalities.
- Section 3 provides an overall synthesis of the developed models and outlines the next steps.

1 Description of available data and splits

The SMASH-HCM project collects two databases from HCM subjects, which are the fundamental sources for data-driven models' development, validation and testing during the project. For the sake of readiness of this document, a brief description of those data sources is provided in the following. The section also contains the criteria adopted to regularize access to the data during the model training/validation/testing phase to avoid bias and overfitting and make the comparison of performance easier and fair.

1.1 SHARE Dataset

1.1.1 Origin of the Collection

The dataset is derived from a multicontinental, multicentre cohort, encompassing patient data from eight centres across Europe, North America, and South America. The data originates from the SHaRe (Sarcomeric Human Cardiomyopathy Registry) dataset, initially comprising approximately 17,000 patients. From this parent dataset, a focused HCM (Hypertrophic Cardiomyopathy) subset of 5,659 individuals was identified. Applying clinical inclusion criteria, 4,591 patients with at least one clinic visit and one echocardiographic measurement of left ventricular (LV) wall thickness were retained. Further filtering based on genetic testing availability resulted in a final analytical cohort of 2,763 patients, which is what has been made available to SMASH-HCM project. There exists significant potential to expand this dataset, by over 60%, by including patients without genetic testing data. This could enable the development of risk stratification algorithms that do not rely on costly genetic testing, offering a more accessible solution for resource-constrained settings and enhancing model generalizability and robustness.

1.1.2 Data Type and Format

The dataset comprises derived clinical measurements extracted from electronic medical records and echocardiography assessments. The data covers a time span of 20 years, with some patients having been followed up for more than 12 years. While it contains comprehensive clinical measurements, only a limited subset of the corresponding raw data is available. Specifically, raw instrument data for echocardiography is not provided, and although digitized ECG traces and CMR images are currently being curated, they are not available for all timepoints. All data underwent HIPAA-compliant pseudonymization at the source, ensuring patient confidentiality and secure handling across international data transfer protocols.

1.2 Rennes Dataset

The Rennes data consists of clinical, genetic, laboratory, medication, procedural, and outcome variables collected from hypertrophic cardiomyopathy (HCM) patients at CHU Rennes, encompassing demographic information (e.g., patient IDs, birth dates, gender), cardiovascular phenotyping (echocardiography, MRI, NYHA classification, LV dimensions, LGE extent, GLS measures), genetic testing results, family and personal cardiac history, comorbidities, biomarkers (e.g., NTproBNP, troponin), exercise testing, prescribed medications, procedural interventions (e.g., ICD implantation, septal myectomy, transplantation), and mortality data. It also includes linked raw imaging and physiological data including baseline ETT and CMR scans, 12-lead ECGs, and raw ETT scans at one-year follow-up.

1.2.1 Origin of the Collection

This dataset comprises retrospectively enrolled adult patients diagnosed with HCM at the University Hospital of Rennes (France) between 2008 and 2020. Patients were included based on current

diagnostic guidelines, following systematic evaluation at a dedicated cardiomyopathy referral center. Exclusion criteria included incomplete or poor-quality echocardiographic data, history of acute coronary syndrome, and significant coronary artery disease. From this population, a subset of 201 patients was selected for analysis, all of whom had complete echocardiographic and 2D speckle-tracking strain data. The mean follow-up duration for this cohort is approximately 4 years, allowing for longitudinal outcome assessment. The study was conducted in accordance with the Declaration of Helsinki and received ethical approval from the institutional review board. Informed consent was obtained from all participants.

1.2.2 Data Type and Format

The dataset includes clinical data extracted from electronic medical records, echocardiographic measurements, including both conventional and speckle-tracking strain features, and raw strain curves. In addition, a subset of patients underwent echocardiographic stress testing, from which selected features were extracted to complement resting assessments. Raw 12-lead ECGs were digitalized from PDF vectorized records, allowing full-feature extraction. Additionally, a subset of patients has MRI-derived variables and genetic testing results, though raw imaging data are not available. All data were pseudonymized in compliance with applicable data protection standards. This cohort serves as a complementary dataset to the SHARE population, enabling multimodal modeling and potential external validation.

1.3 A unified approach to data analysis

In order to harmonize the efforts of model developers and to make model results comparable, a series of decisions have been taken on the approach to the available data:

- Three main endpoints were identified, and models will be trained for those endpoints. They are:
    - *Composite_HF* (first occurrence of cardiac transplantation, LV assist device implantation, LV ejection fraction <35%, or New York Heart Association class III/IV symptoms NYHA class III-IV);
    - *Composite_VArrhythmia* (first occurrence of sudden cardiac death, resuscitated cardiac arrest, or appropriate implantable cardioverter-defibrillator therapy SCD, aborted cardiac arrest, or appropriate ICD therapy);
    - *Composite_Overall* (irst occurrence of any component of the ventricular arrhythmic or heart failure composite end point (without inclusion of LV ejection fraction), all-cause mortality, atrial fibrillation (AF), stroke, or death).
- A shared vision on how to approach event prediction and/or progression of disease has been decided, and models will be trained for predicting Time-to-Event (where Time=Time of Visit, Event=Endpoint time) or the endpoint-risk (i.e., risk that an endpoint will occur within 5 years from the Time of Visit)
- The amount of contextual data used for prediction/inference must be indicated and referred to the Time of event.
- Data split has been unified and shared among all the partners. In accordance with WP2, a *csv* file has been created and uploaded to the databank. The file indicates if patients' ID belongs to train (and which fold) or to test split. Splits have been created on a patient base, and they have been stratified so to guarantee the maximum of homogeneity among splits and folds. The rigid division will avoid abuse of data during training and facilitate comparison of the performances.

While those constraints may sound restrictive, they will tend to harmonize the approach to the data, will focus model development on specific targets of clinical relevance, will allow immediate comparison of the performances and potentially limit overfitting issues. The harmonization will also facilitate combination and integration of models in task T6.4.

2 Descriptions of Initial data-driven models and results

This section introduces the initial set of data-driven models developed within the SMASH-HCM project. The presentation is structured as follows: first, we describe models based on a single modality (ECG, echocardiography, or CMR), followed by models that integrate multiple modalities or parameter combinations. For each model, preliminary results obtained on the SHARE and/or Rennes datasets are also provided.

2.1 ECG-based Models

We adopted two approaches for the development novel ECG biomarkers and the relative data-driven models. We developed: a) feature-based ECG model and b) end-to-end ECG model.

2.1.1 Feature-based ECG models

These models are based on the computation of a series of comprehensive ECG features. Three categories of features were systematically extracted from the ECG signals (and a dedicated tool was developed) to capture different aspects of electrical and structural cardiac abnormalities (see also Figure 1):

1. Morphological Features: We computed 24 features per lead, including intervals (PR, PS, PT, QT, QRS, RS, ToT, and TTe (To: T offset, Te: T end)), fiducial point amplitudes (P, Q, R, S, J, T), slope measurements (QR, RS, SJ, T-wave), amplitude ratios (R/P, R/T), percentage of negative QRS area, and slope ascending and descending (T and P waves and QR, RS and SJ intervals). Additional global features included the QRS and T axis deviations and frontal QRS angle. These were derived using the ECGDeli[1] toolbox available in MATLAB.
2. Hermite Transform Coefficients: The average QRS complex per lead was approximated using the first four Hermite basis functions[2], providing a compact representation of QRS morphology. The Root Mean Square error (RMSE) between the original QRS and its Hermite approximation was also computed to assess morphological variability.
3. Variational Mode Decomposition (VMD): Each lead was decomposed into five intrinsic mode functions. For selected modes (u_3, u_4, u_5), we extracted the central frequency and the number of local extrema (optima) per QRS complex. The total number of extrema in the original QRS complex was also included as a feature.

A total of 436 features were generated per patient, encompassing all leads and methods (Table 1).

Table 1: Summary of the biomarkers with their numbers.

Name	For each lead	number
Median fiducial point amplitude (P, Q, R, S, J, T)	Yes	6x12= 72
Ratio of median fiducial amplitudes (R/P, R/T)	Yes	2x12=24
Median of the interval length (PR, PS, PT, QT, QRS, RS, ToT, TTe)	Yes	8x12=96
Slope (T asc, T des, P asc, P des, QR, RS, SJ)	Yes	7x12=84
Negative percentage of QRS	Yes	1x12=12
Median RR	No	1
QRS axis deviation, T deviation, and frontal QRS angle	No	3
Hermite coefficient	Yes	4x12=48
RMSE between Hermite approximation and the original QRS	Yes	1x12=12
VMD (# of local optima + central frequency in 3 modes, + local optima in QRS)	Yes	7x12=84
Total		436

[1] Pilia N. et al. *ECGdeli - An open source ECG delineation toolbox for MATLAB*, SoftwareX, vol. 13, no. 100639, 2021.

[2] Sornmo L. et al. *A Method for Evaluation of QRS Shape Features Using a Mathematical Model for the ECG*, IEEE TBME, pp 713-717, 1981.

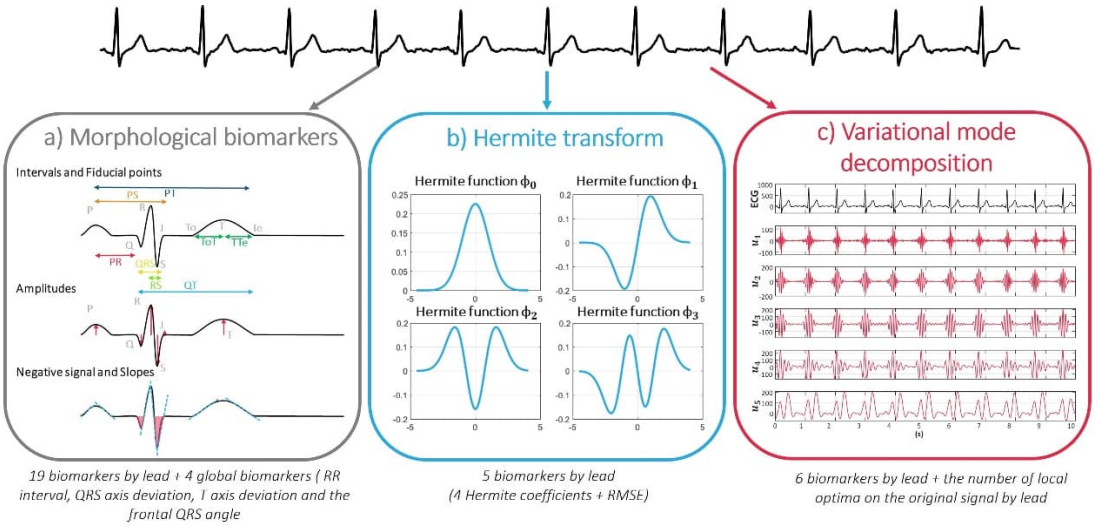

Figure 1: a) Morphological features extraction: Interval width, negative % of the signal, amplitude of the fiducial points and slopes of the waves. (To: T offset, Te: T end) b) The first four Hermite functions ($\mathbf{\Phi_0}$, $\mathbf{\Phi_1}$, $\mathbf{\Phi_2}$ and $\mathbf{\Phi_3}$). c) Variational mode decomposition with the original ECG signal and the 5 modes ($\mathbf{u_k}$).).

Arrhythmic Events Risk Model

The futures were analysed for predicting risk of arrhythmic composite events in the Rennes cohort of HCM patients. HCM patients were divided into 2 groups: 1) HCM patients with higher risk of arrhythmic events vs. 2) HCM control patients. The Composite_VArrhythmia endpoint was used, and a patient was assigned to the higher risk class if one of the events of the Composite_VArrhythmia endpoint happened during a mean follow-up period of 1070 days.

We analyzed 12-lead ECGs recorded over 10 seconds from the first available patients of the Rennes University Hospital cohort (10s, 12 leads synchronized ECG from 55 patients diagnosed with HCM). The original ECGs were extracted from vectorized PDF files using a custom Python pipeline (Figure 2), which enabled the conversion of graphical traces into digital signals. These signals were subsequently resampled at 1 kHz to allow for precise detection and quantification of ECG features. All recordings were originally acquired with synchronized leads and a sampling rate of 250 Hz.

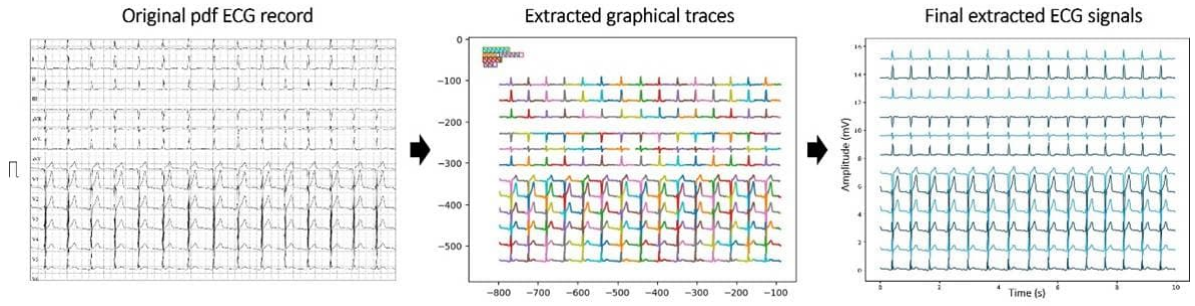

Figure 2: Extraction of the vectorized ECG records.

To identify features most relevant to arrhythmic risk, we applied a voting-based feature selection strategy that combined four univariate ranking metrics: Fisher score, Kullback-Leibler divergence, Mann-Whitney U test, and AUC. Features ranking consistently across multiple metrics were selected for further analysis.

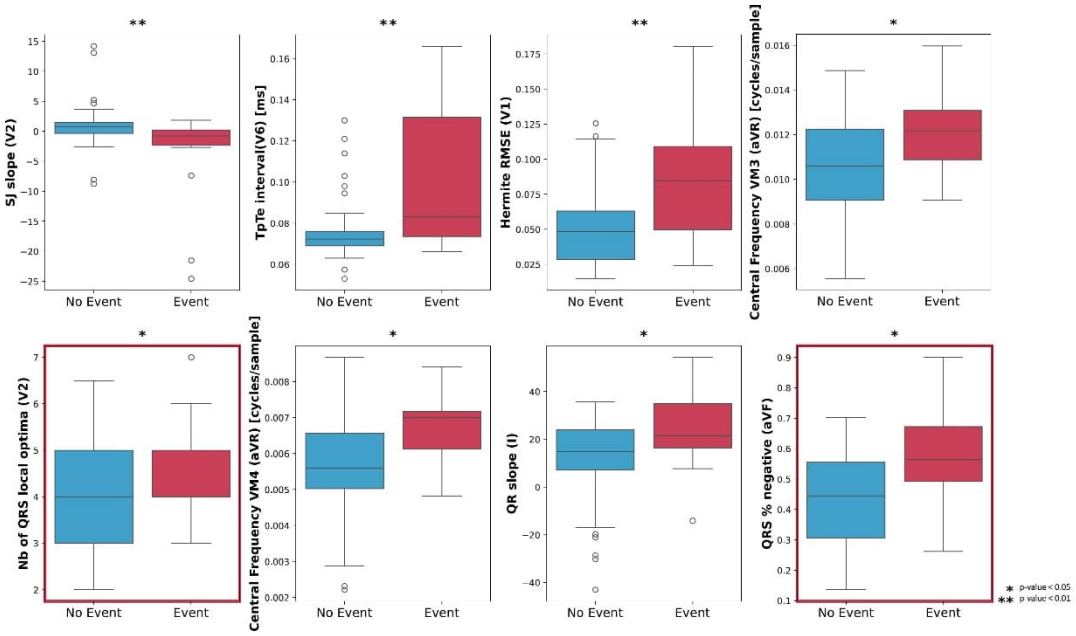

Figure 3: Boxplot of event vs non-event patients of the SJ slope (V2), TpTe interval (V6), Hermite RMSE (V1), centrale frequency VM3 (aVR), number of QRS local optima (V2), central frequency VM4 (aVR), QR slope (I), QRS percentage negative (aVF) with the 10 lower U test p-value (∗ < 0.05, ∗∗ < 0.01).

Among the 8 selected features presented in Figure, the number of QRS local optima in lead V2 and the percentage of negative QRS area in lead aVF were most predictive. Kaplan-Meier survival analysis (Figure 4) revealed significant differences in event-free survival between groups stratified by the median values of these features ($p < 0.01$ and $p < 0.05$, respectively). Entropy analysis further demonstrated that low-risk groups showed lower internal variability, supporting the robustness of these biomarkers for patient stratification. Results have been presented at IEEE-EMBS Conference 2025[3].

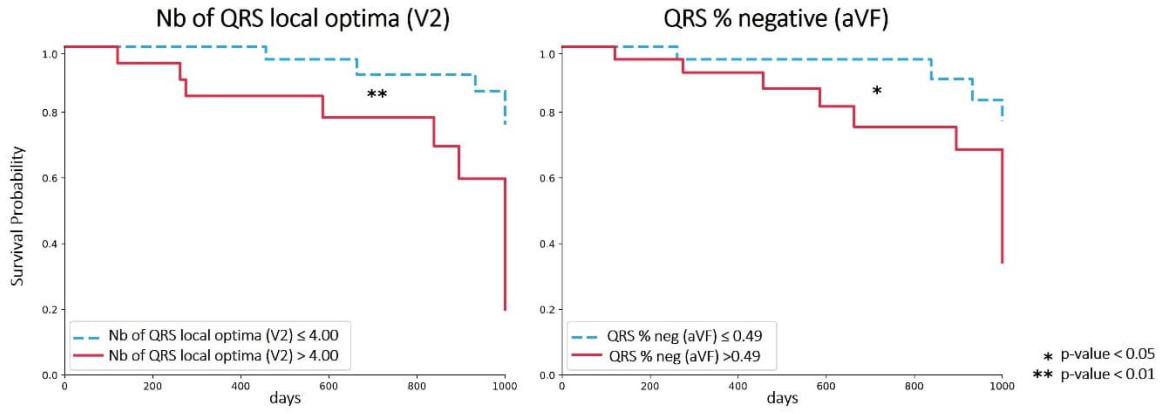

Figure 4: Survival curves of the two biomarkers presented before with the lowest LogRank test p-value (∗ < 0.05, ∗∗ < 0.01): number of QRS local optima (V2) and QRS percentage of negative (aVF).

[3] Taconné M., Corino V., Melot A., Al Wazzan A., Donal E., Cerveri P., Mainardi L., *Extraction of Risk Markers from ECG in Patients with Hypertrophic Cardiomyopathy*. IEEE EMBC 2025.

Hypertrophy classifiers

To further evaluate the potential of ECG-derived features in detecting structural abnormalities (early signs of HCM), we developed and validated classifiers for left and right ventricular hypertrophy (LVH and RVH). These models were trained and tested using publicly available datasets.

RVH Classifier. The RVH classifier was trained on the PTB-XL database (101 RVH, 8900 controls) and validated on the Georgia ECG Challenge database (69 RVH, 1387 controls). Features were extracted from the ECGs, using the methodology described in the previous section. Feature selection using sequential floating forward selection (SFFS), applied on the training part of the cross-validation on the PTB-XL database, led to a compact set of features (< 30 features per fold). Among the three tested classifiers: logistic regression, random forest, and support vector machines, the R amplitude in lead V1 consistently ranked first in all models. Other top features included the Hermite RMSE in lead II or aVR, the descending slope of the T wave, and, for SVM, the S amplitude in lead I. These features align with clinical expectations: tall R waves in right precordial leads and abnormal repolarization are known indicators of RVH. The low number of selected features enhances the model's interpretability and generalizability. After the cross-validation applied on the PTB-XL database an external validation on the Georgia cohort was applied and showed excellent results in Table 2. Despite the SVM model achieving the highest accuracy and specificity, its poor sensitivity limits its clinical utility for RVH screening. The RF and logistic regression models, however, demonstrated an excellent balance between specificity and sensitivity. Overall, these results confirm that ECG-derived features, even in limited number, can robustly identify RVH in external populations. The complete study was presented at Computing in Cardiology 2024[4].

Table 2: Sensitivity, specificity, and accuracy of the three model on the validation database: Logistic Regression, Random Forest and Support Vector Machine.

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUROC
Random Forest	88.0 ± 3.6	87.0 ± 3.1	88.1 ± 3.8	0.93±0.01
Logistic Regression	95.9 ± 0.7	84.5 ± 1.3	96.5 ± 0.7	0.91±0.002
Support Vector Machine	97.0 ± 0.0	58.0 ± 0.0	99.0 ± 0.0	0.91±0.0001

LVH Classifier. For LVH, we trained models on the PTB-XL database (2181 LVH, 9001 controls) and validated them on the Georgia database (1012 LVH, 1387 controls). A consistent ECG feature set was extracted using the same pipeline as for RVH, and feature selection was again performed using SFFS for each classifier. The performances are summarized in Table 3 and the ROC curves are displayed in Error! Reference source not found. with the best five clinical criteria: Siegel[5], McPhie[6], Sokolow-Lyon[7], Gant[8], Romhilt[9], as comparison to the state of the art. Ensemble models were created based on the 100 models trained during the cross-validation on the PTB-XL database. Random forest and SVM classifiers performed best on the external Georgia database, with AUCs exceeding 0.97 on the external cohort. Even simplified versions of the models, using only the top five selected features (e.g., RF5, SVM5), maintained strong performance (e.g., RF5: AUC = 0.96). The features most often selected

[4] Taconné M. Corino V., Mainardi L., *Automatic Right Ventricular Hypertrophic Detection Integrating Electrocardiography-based QRS Biomarkers with Machine Learning*. CinC 2024

[5] R. J. Siegel and W. C. Roberts, *Electrocardiographic observations in severe aortic valve stenosis: Correlative necropsy study to clinical, hemodynamic, and ECG variables demonstrating relation of 12-lead QRS amplitude to peak systolic transaortic pressure gradient,* Amer. Heart J., vol. 103, no. 2, pp. 210–221, 1982

[6] J. McPhie, *Left ventricular hypertrophy: Electrocardiographic diagnosis,* in Australas. Ann. Med., vol. 7, no. 4, pp. 317–327, 1958.

[7] M. Sokolow and T. P. Lyon, *The ventricular complex in left ventricular hypertrophy as obtained by unipolar precordial and limb leads*, Amer. J. Med., vol. 37, no. 2, pp. 161–186, 1949.

[8] R. Grant, *Clinical Electrocardiography: The Spatial Vector Approach.* New York, NY, USA: McGraw-Hill Blakiston Division, 1957.

[9] D. W. Romhilt and E. H. Estes, *A point-score system for the ECG diagnosis of left ventricular hypertrophy*, Amer. Heart J., vol. 75, no. 6, pp. 752–758, 1968.

included R amplitudes in leads V1, V5, V6, T wave slopes, and Hermite RMSE, which are also known to correlate with classical ECG criteria for LVH. When compared to traditional clinical criteria such as Sokolow-Lyon, Romhilt-Estes, or Siegel scores, the machine learning models exhibited significantly higher sensitivity while maintaining high specificity. The combination of five clinical criteria reached a sensitivity of 76% and specificity of 95% but still fell short of the best-performing ML models. The complete study was published in IEEE OJEMB[10].

Table 3: Sensitivity, specificity, accuracy and balanced accuracy of the ensemble method for the 6 types of model on the validation database: Logistic Regression (LogReg), Random Forest (RF) and Support Vector Machine (SVM) for the selected features up to 30 features and to the reduced to 5 features. The sensitivity, specificity, and accuracy of the 5 best literature/clinical criteria were also added.

Model / Criterion	Sensitivity	Specificity	Accuracy	Balanced Accuracy	AUROC
Ensemble LogReg	0.724	0.941	0.850	0.833	0.926
Ensemble LogReg5	0.723	0.909	0.831	0.816	0.882
Ensemble RF	0.905	0.933	0.921	0.919	0.976
Ensemble RF5	0.875	0.925	0.904	0.900	0.959
Ensemble SVM	0.875	0.955	0.922	0.915	0.978
Ensemble SVM5	0.868	0.938	0.908	0.903	0.970
Siegel	0.467	0.960	0.752	0.714	--
McPhie	0.275	0.993	0.690	0.634	--
Sokolow-Lyon1	0.384	0.997	0.739	0.691	--
Gant2	0.124	0.997	0.629	0.561	--
Romhilt1	0.148	0.984	0.632	0.566	--
Combination of 5 clinical criteria	0.762	0.954	0.873	0.858	--

Since left ventricular hypertrophy is a hallmark sign of HCM, we aimed to develop ECG-based classifiers capable of identifying LVH and RVH accurately. Beyond offering robust classification tools, these models confirm that the extracted ECG features are sensitive to the presence and potentially the localization of ventricular hypertrophy. This supports their relevance for HCM monitoring and risk prediction and motivates further investigation into their integration in broader multi-modal risk models. If ECG-derived features can reliably quantify or localize hypertrophy, we hypnotize that they may also contribute to individualized risk stratification and management in patients with HCM.

[10] Taconné M., Corino V., Mainardi L., *An ECG-Based Model for Left Ventricular Hypertrophy Detection: A Machine Learning Approach.* IEEE OJEMB, vol.6 p219-226, 2025. DOI: 10.1109/OJEMB.2024.3509379

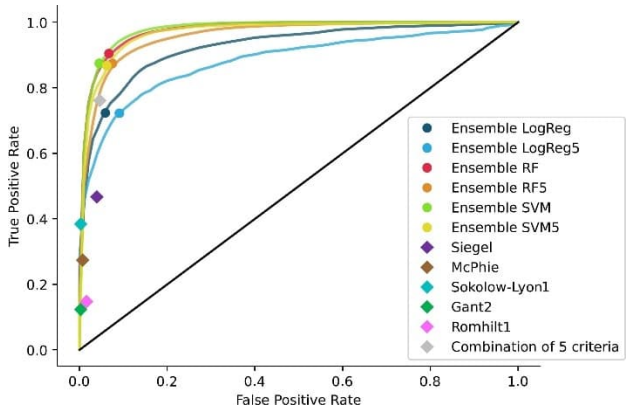

Figure 56: Mean ROC curves of the 3 algorithms on the external validation database: Logistic Regression (LogReg), Random Forest (RF) and Support Vector Machine (SVM) for the selected features up to 30 features and to the reduced to 5 features: LogReg5, RF5, SVM5, with the point of their equivalent ensemble method. The best 5 clinical criteria: Siegel[11], McPhie[12], Sokolow-Lyon[13], Gant[14], Romhilt[15] and their combination are also plotted, on the external validation database.

2.1.2 End-to-End models

The original plan was to derive deep (CNNs/RCNNs, Autoencoders), transformers architectures to derive novel ECG-based biomarkers. However, the paucity of digitalized ECG traces, in the SHARE and in the Rennes datasets, has prevented the investigation of end-to-end architecture (CNN, Autoencoders, transformers) and the evaluation of their performances on SMASH-HCM data. While sufficient digitalized ECGs are going to be collected and curated, we identified a few networks, designed for end-to-end ECG analysis, whose architecture were trained and parameters are publicly available, and we investigated their performances for risk stratification of SMASH-HCM patients.

HCM Autoencoder

A pretrained variational autoencoder (VAE) was made available by the following study[16]. This model was trained on a private dataset of subjects with hypertrophic cardiomyopathy (HCM). The VAE takes 12-lead ECG signals as input, specifically, the median beat from each lead and encodes them into 24 latent features. The pretrained VAE was applied to a subset of 12-lead ECG signals from the Rennes dataset, for which a synchronized median beat could be extracted using the code provided by the authors. No additional processing was needed, as the code provided by the authors was already sufficiently detailed and self-contained. For now, the analysis is limited to 124 ECG recordings (the ones that were digitalized at the time of this Deliverable). The 24 latent features were then used to train a set of ML classifiers to predict the *Composite_VArrhythmia* outcome. Among the 124 cases, only 10 met this endpoint, highlighting a substantial class imbalance.

Statistical analysis: Despite the limited number of events, a preliminary statistical analysis was performed to investigate the association between the 24 VAE-derived features and the arrhythmic

[11] R. J. Siegel and W. C. Roberts, *Electrocardiographic observations in severe aortic valve stenosis: Correlative necropsy study to clinical, hemodynamic, and ECG variables demonstrating relation of 12-lead QRS amplitude to peak systolic transaortic pressure gradient,* Amer. Heart J., vol. 103, no. 2, pp. 210–221, 1982

[12] J. McPhie, *Left ventricular hypertrophy: Electrocardiographic diagnosis,* in Australas. Ann. Med., vol. 7, no. 4, pp. 317–327, 1958.

[13] M. Sokolow and T. P. Lyon, *The ventricular complex in left ventricular hypertrophy as obtained by unipolar precordial and limb leads*, Amer. J. Med., vol. 37, no. 2, pp. 161–186, 1949.

[14] R. Grant, *Clinical Electrocardiography: The Spatial Vector Approach.* New York, NY, USA: McGraw-Hill Blakiston Division, 1957.

[15] D. W. Romhilt and E. H. Estes, *A point-score system for the ECG diagnosis of left ventricular hypertrophy*, Amer. Heart J., vol. 75, no. 6, pp. 752–758, 1968.

[16] Carrick R. et al. Identification of high-risk imaging features in hypertrophic cardiomyopathy using electrocardiography: A deep-learning approach, Heart Rhythm, vol. 21, no. 8, pp 1390-1397, 2024. DOI: 10.1016/j.hrthm.2024.01.031

outcome. A non-parametric Mann–Whitney U test was used with a significance level of 0.05. Four features were found to be statistically significant and are shown in Fig. 8.

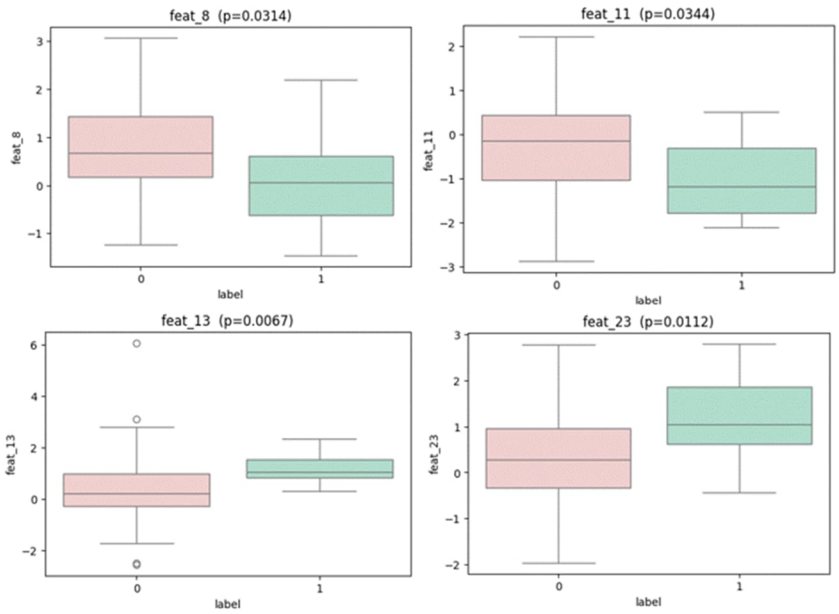

Figure 7: Distribution of the statistically significant features identified in the analysis

Machine learning: Different ML strategies were explored in an attempt to train a reliable model despite the limited number of positive cases. The best-performing approach had the following characteristics. A nested cross-validation strategy was adopted, with 5 folds in the outer loop (used for model evaluation) and 5 folds in the inner loop (used for model selection). Within the inner loop, a grid search was conducted to optimize hyperparameters. To address the class imbalance, a two-step resampling strategy was applied only on the training data in each inner fold. First, the majority class was downsampled to three times the size of the minority class. Then, SMOTE was applied to oversample the minority class and balance the training set. Across all five outer folds, the best-performing model was an SVM, with a different optimal hyperparameter configuration selected by the inner cross validation. Other classifiers, including decision trees, logistic regression, naive Bayes, and k-nearest neighbors, were also tested but consistently yielded inferior results; for this reason, their outcomes are not reported here. For completeness, the ROC curves corresponding to the five outer folds using the SVM classifier are shown in Figure 8.

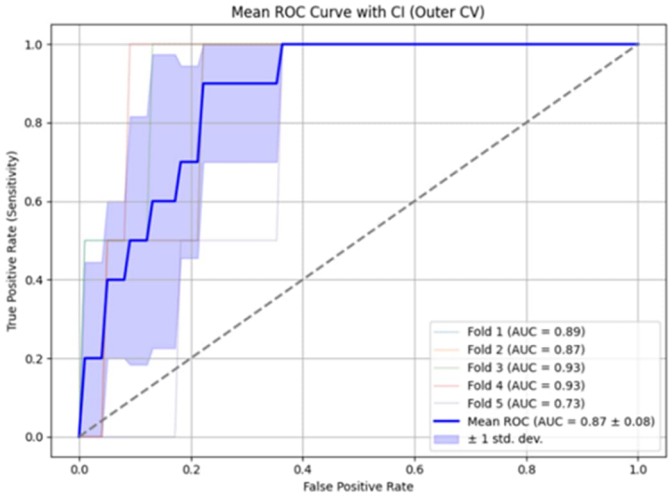

Figure 8. ROC curves for the five folds of the outer cross-validation

2.2 CMR-based models

Within this section, we now focus on the Cardiac Magnetic Resonance (CMR) imaging processing and analysis. Our work advances cardiac MRI analysis in three ways. We developed deep learning–based reconstruction methods that improve image quality and efficiency, a harmonic map framework that enables clearer visualization and biomarker discovery from cardiac simulations, and synthetic MRI generation techniques to balance underrepresented disease categories.

2.2.1 CMR Reconstruction

One of our contributions has been the development of CMR reconstruction methods based on unrolled models. These approaches aim to improve the quality of cardiac MRI images while reducing acquisition time, ensuring more efficient use of available data. The methodology builds upon recent advances in model-based deep learning, where the reconstruction pipeline is reformulated into a trainable architecture. This allows us to better capture the complex structure of cardiac data and to enhance the robustness of the reconstructed images. Current work has already demonstrated promising preliminary results, and further optimization and validation are ongoing.

2.2.2 Visualization and Biomarker Discovery Using Harmonic Maps

We are also developing a framework for visualization and biomarker discovery using harmonic maps applied to cardiac action potential (AP) simulations. This approach provides novel ways to interpret high-dimensional simulation data, enabling the identification of meaningful patterns and potential biomarkers associated with disease phenotypes. By mapping AP dynamics into harmonic spaces, we can generate interpretable visual representations that support both exploratory data analysis and clinical insight. Ongoing work focuses on refining these methods and validating them against available annotated datasets.

The proposed framework employs disk harmonic mapping flattening techniques to transform the left ventricle into flattened two-dimensional representations, enabling enhanced static visualization and extending to dynamic spatiotemporal analysis of cardiac electrical activity, introducing a novel standardized approach for electrophysiological pattern analysis of spiral waves using the American heart's association bull's eye plot. As a result, this study can improve diagnostic accuracy in clinical settings and provide new insights into cardiac arrhythmia patterns.

2.2.3 Cardiac MRI Synthesis for Dataset Balancing

A third line of contribution (see[17] for more details) involves cardiac MRI synthesis techniques designed to balance datasets across different disease categories. Since many cardiovascular conditions are underrepresented in existing collections, synthetic generation of CMR data provides an opportunity to mitigate class imbalance and improve the performance of downstream machine learning models. Our approach leverages generative methods to create realistic disease-specific variations, thereby strengthening the diversity and representativeness of the databank. We found that using synthetic scans to balance the dataset resulted in up to a 0.05 increase in the DICE score.

Our proposed method combines modifications of an existing atlas that represent a healthy patient with image-to-image style transfer to produce synthetic subsets of data that constitute additional images to train segmentation networks. The mentioned method was performed over publicly available data.

The proposed method is founded on the premise that inaccuracies in model predictions stem from the unique cardiac morphology observed in patients with pathologies. Our hypothesis is that the absence

[17] : https://link.springer.com/chapter/10.1007/978-3-031-66958-3_26

of patients with specific pathologies in the training data can cause the model to overfit the pathologies that are present. To mitigate this, we introduce scans that exhibit morphology like the pathologies not originally represented in the training data. This augmentation increases the dataset's diversity, resulting in enhanced segmentation model performance. The overall structure of the method's pipeline is depicted in Figure 9.

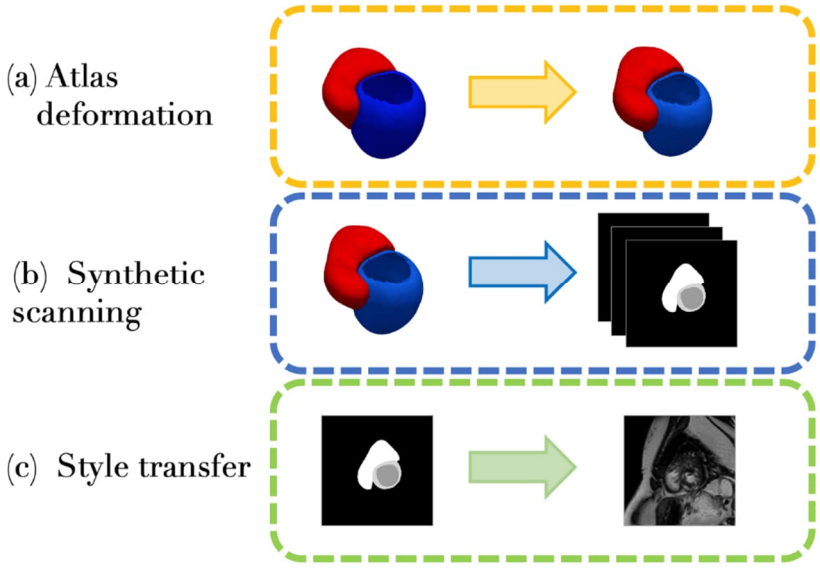

Figure 9: The pipeline of the image synthesis method. The three main components correspond to: (a) the atlas deformation part, (b) the synthetic scanning component, and (c) the style transfer part.

Table 4. Segmentation performances for each subset of data. Each subset presents the DICE score (higher is better) and Hausdorff distance (lower is better) for each of the four models and each of the regions (Left Ventricle, Myocardium, and Right Ventricle). Models refer to the model trained with healthy patients of the original dataset plus the atlas or its synthetic deformed aliases.

Data subset	Model	DICE				Hausdorff (mm)
		LV	MYO	RV	ALL	All
Normal (Healthy)	Healthy	0.942	0.825	0.854	0.874	6.311
	HCM	0.940	0.840	0.827	0.869	6.292
	DRV	0.933	0.838	0.805	0.859	6.488
	Augs	0.948	0.842	0.879	0.889	6.422
HCM	Healthy	0.938	0.858	0.828	0.875	6.554
	HCM	0.919	0.857	0.832	0.870	6.608
	DRV	0.915	0.845	0.824	0.861	6.719
	Augs	0.909	0.820	0.811	0.847	6.635
DRV	Healthy	0.781	0.604	0.774	0.719	7.346
	HCM	0.762	0.615	0.766	0.715	7.141
	DRV	0.805	0.650	0.772	0.742	6.902
	Augs	0.792	0.624	0.714	0.710	6.503
All	Healthy	0.858	0.715	0.772	0.781	6.817
	HCM	0.831	0.714	0.786	0.777	6.770
	DRV	0.867	0.739	0.773	0.793	6.903
	Augs	0.858	0.723	0.767	0.782	6.547

To rectify the imbalance in our training data, we introduced what we refer to as "synthetic patients." These synthetic patients are created by applying deformations to an atlas, essentially simulating the heart structures of the two previously mentioned diseases, Hypertrophic Cardiomyopathy (HCM) and Dilated Right Ventricle (DRV). These simulations are achieved by virtually slicing the 3D model of the heart to generate these synthetic patient representations.

The incorporation of these synthetic patients into the dataset involves a crucial style transfer phase. During this step, images that resemble labels or structural representations are transformed into realistic MRI scans. We accomplished this transformation using a technique called Contrastive Unpaired Translation.

Finally, a segmentation network is trained using real and synthetic data. The results in Table 4 show how the addition of synthetic data had a positive impact on the performance, with a strong emphasis on the dilated right ventricle deformations. Overall performance increased the DICE score in that model by 0.01, and by 0.05 in the myocardium for the DRV subset of data, where the mean performance was more than 0.02 better than in the healthy subset.

In our study, we successfully trained a style transfer model to effectively generate synthetic cardiac MRI images that are sampled from a deformed atlas. Those synthetic images were successfully used to balance the training dataset. The results show how the addition of this data meant an increase in the DICE score of up to 0.05 in some regions within the target pathologies of the data.

2.3Echocardiographic Models

As a first step toward risk prediction in HCM, we developed a machine learning-based risk model relying solely on conventional echocardiographic features. The study was based on 1,061 HCM patients of the SHARE registry, all followed at Careggi University Hospital (Florence). Each patient had a baseline echocardiographic exam, with features routinely collected in clinical settings. The outcome was defined as a composite 5-year endpoint that included both progression to heart failure (e.g., NYHA class III-IV, LVEF < 35%, device implantation, or cardiac transplant) and major arrhythmic events (e.g., SCD, aborted cardiac arrest, or appropriate ICD therapy).

A total of 34 echocardiographic parameters were included, covering structural, functional, and hemodynamic features: left ventricular dimensions and wall thicknesses, left atrial volume, LV ejection fraction (LVEF), diastolic indices (E/E'), presence and grade of valve regurgitation or stenosis, and left ventricular outflow tract (LVOT) gradients at rest and with provocation. After removing highly correlated variables based on clinical knowledge (correlation threshold > 0.75), 26 features were retained for modelling.

Four machine learning classifiers: logistic regression, support vector machines (SVM), random forest, and gradient boosting, were implemented within a nested cross-validation pipeline to avoid overfitting and ensure generalizable performance estimates. Each model included preprocessing steps such as median imputation, standardization, and random under-sampling to mitigate class imbalance.

A summary of the cross-validated performance for each model is presented in Table 5. To visualize discriminative ability, Figure 10 Error! Reference source not found.shows the mean ROC curves for the four models across the five outer folds, along with the mean ROC curve over the same folds of the ESC risk score for comparison. Among all models tested, logistic regression achieved the best trade-off between sensitivity and specificity, with a balanced accuracy of 73.6%, sensitivity of 72.2%, and specificity of 75.0%. Its simplicity and interpretability make it particularly attractive for clinical applications.

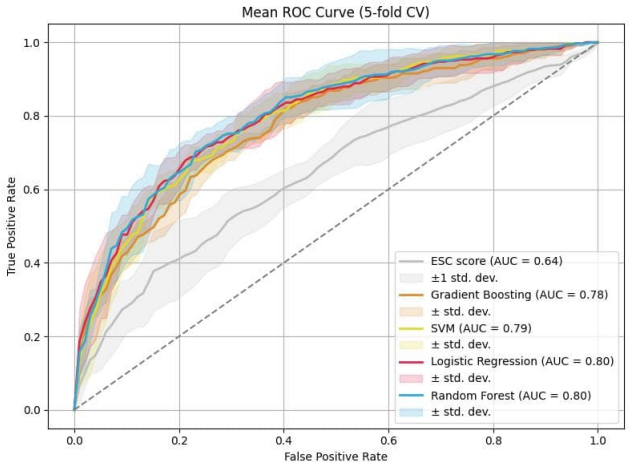

Figure 10: ROC curves of the 5-CV models with area under the curve (AUC), compared to the ESC risk score.

Feature importance analysis, based on logistic regression coefficients and SHAP values for three other models, consistently identified key predictors aligned with clinical knowledge: left ventricular diameter, maximum wall thickness (MWT), LVOT obstruction gradient and location, age, and LVEF. Additional important variables were mitral and tricuspid valve regurgitation, posterior wall thickness, and velocities, indicators of diastolic dysfunction and disease progression.

Table 5: Cross-validated performance metrics (%) for each classifier. Values are reported as mean ± standard deviation across the five outer folds.

Metric	Random Forest	Logistic Regression	SVM	Gradient Boosting
Sensitivity	73.4 ± 6.7	72.2 ± 4.4	68.4 ± 3.5	70.1 ± 4.7
Specificity	72.6 ± 4.6	75.0 ± 3.5	75.2 ± 3.5	70.7 ± 4.5
Balanced Accuracy	73.0 ± 2.6	73.6 ± 1.2	71.8 ± 1.5	70.4 ± 1.6
F1 Score	63.5 ± 3.0	64.2 ± 1.5	62.1 ± 1.8	60.5 ± 2.0

Longitudinal analysis

To explore the value of echocardiographic modeling for patient follow-up, we analyzed risk trajectories over time in a subset of patients who underwent serial echocardiographic exams. The model trained on the first outer fold of the nested cross-validation was retrospectively applied to all earlier exams available for each patient in the corresponding test set, generating a time series of predicted risk scores.

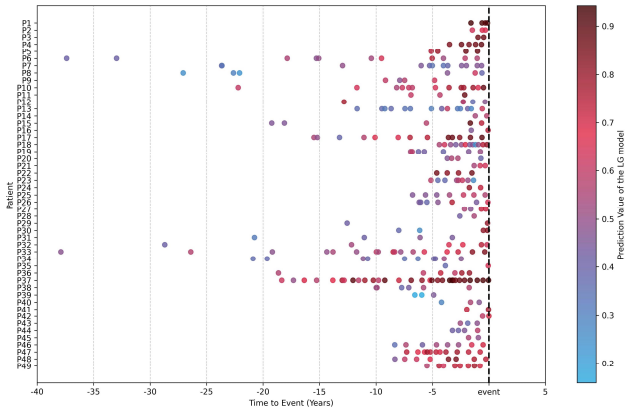

Figure 11: Predicted risk of the best model through follow-up of patients (P1-P49) with event in the first test set fold.

Out of the patients who experienced the composite endpoint, 49 had at least two echocardiographic follow-up exams and were included in the longitudinal analysis. For each of these patients, we computed the evolution of the model-predicted 5-year risk across time. The analysis, illustrated in Figure 11, revealed that in 83.7% of cases, the predicted risk increased progressively prior to the clinical event, with a mean annual rise of 3.0% ± 5.5. This trend suggests that the model is not only useful for static risk prediction at baseline but can also track disease evolution over time.

2.4 Multi-Modality Models

Following the development of the initial echocardiographic risk model, and based on feedback from clinical partners, we extended the feature set by incorporating additional variables derived from clinical examinations and other modalities. These include blood pressure measurements, NYHA functional class, and patient age. These features are routinely collected and available across both clinical sites involved in the study (SHARE and Rennes). To ensure compatibility across cohorts, the original feature set was adjusted to include only variables common to both tabular datasets.

2.4.1 Echocardiographic+Clinical Features

The refined dataset includes 22 features and 1037 patients (few patients were excluded due to missing data), combining echocardiographic and clinical parameters:

- Echocardiographic features:
  LVEDV, LVESV, MiV Regurg, E' septal, LVOTO at Rest, LVOTO at Rest Hypovalue,
  LA diameter, LVEF, LVIDd, LVIDs, Max LVT, Max LVT Location, PW, SAM, E DT, E Septal/ E'
- Clinical features:
  NYHA class, BP Systolic, BP Diastolic, Age

Models were retrained on SHARE patients using only the two best-performing classifiers from the previous analysis: logistic regression and random forest on this harmonized feature set. The results indicated an improvement in predictive performance over the previous models (Table 6).

Table 6: Cross-validated performance metrics (%) for the 2 best classifiers with the refined dataset of Florence and added clinical features. Values are reported as mean ± standard deviation across the five outer folds.

Metric	Logistic Regression	Random Forest
Sensitivity	78.8 ± 4.7	79.4 ± 3.7
Specificity	76.1 ± 5.6	77.2 ± 3.8
Balanced Acc.	77.4 ± 4.8	78.3 ± 3.6
F1 Score	69.0 ± 5.7	70.0 ± 4.4
AUC	86.0 ± 3.0	85.0 ± 3.0

Subsequently, additional variables were incorporated to further expand the multi-modal nature of the model. These included family history (HCM and SCD in the family), comorbidities (Hypertension, Unexplained_Syncope, Chest Pain), arrhythmic indicators (NSVT), and medication data (Beta Blocker, Diuretic, L-Type Calcium Blockers, SGLT2i, RAAS). While the inclusion of these features did not lead to a significant performance gain in global metrics (Table 7), they were nonetheless frequently ranked among the most influential predictors in feature importance analyses. This suggests that they may still carry relevant prognostic information, especially in subgroups or as part of ensemble approaches.

These developments reflect the ongoing effort to transition from single-modality models to integrated risk prediction frameworks that more comprehensively reflect the clinical complexity of HCM.

Table 7: Cross-validated performance metrics (%) for the 2 best classifiers with the refined dataset of Florence and added clinical features, family history, comorbidity and medication. Values are reported as mean ± standard deviation across the five outer folds.

Metric	Logistic Regression	Random Forest
Sensitivity	78.8 ± 5.8	78.5 ± 5.2
Specificity	76.5 ± 4.0	79.6 ± 4.3
Balanced Acc.	77.6 ± 4.3	79.1 ± 4.1
F1 Score	69.1 ± 5.1	71.0 ± 5.0
AUC	86.0 ± 4.0	85.0 ± 3.0

Preliminary validation on Rennes Data

To assess the generalizability of our model, we performed an external validation on the Rennes cohort, which includes 382 HCM patients with available tabular data. The same two classifiers: logistic regression and random forest, were applied using the previously selected set of 22 features. We tested both versions of the feature set: the base model including only echocardiographic and clinical features, and the extended model with added comorbidities and medication data. We retrained the best-performing models on the full SHARE dataset, using optimized hyperparameters derived from the cross-validation.

Table 8: Validation of the two models with the two version of features set on the external Rennes tabular data. The models were retrained on the entire SHARE dataset.

Metric	Logistic Regression *ECHO + clinical*	Random Forest *ECHO + clinical*	Logistic Regression *ECHO + clinical +comorbidities + medication*	Random Forest *ECHO + clinical +comorbidities + medication*
Sensitivity	68.4	70.2	68.4	63.2
Specificity	66.8	60.3	67.4	67.7
Bal. Acc.	67.6	65.2	67.9	65.4
F1 Score	38.2	35.4	38.6	36.4
AUC	70.9	70.7	70.9	70.1

A marked drop in performance was observed across all metrics when applied directly to the Rennes dataset, highlighting a domain shift between cohorts (Table 8). The drop persisted regardless of whether the added variables (e.g., family history, medications) were included. These results indicate that even with harmonized features, site-specific factors such as acquisition protocol differences, data quality, or clinical annotation practices can significantly affect model performance.

Additionally, we implemented an ensemble strategy based on the five logistic regression models trained during cross-validation. A majority voting scheme was applied across their predictions to create a consensus output. This ensemble approach (Table 9) yielded slightly improved results compared to the single retrained model, offering increased robustness across folds and marginal gains in predictive performance.

Table 9: Validation of the two models with the two version of features set on the external Rennes tabular data. The models are the ensemble model of the 5 trained models of the cross-validation previously presented.

Metric	Logistic Regression *ECHO + clinical*	Random Forest *ECHO + clinical*	Logistic Regression *ECHO + clinical +comorbidities + medication*	Random Forest *ECHO + clinical +comorbidities + medication*
Sensitivity	66.7	70.2	71.9	66.7
Specificity	67.4	60.9	64.6	68.3
Bal. Acc.	67.0	65.5	68.3	67.5
F1 Score	37.8	35.7	38.5	38.4
AUC	72.8	72.4	72.7	70.9

At the end of the validation process, we compared the ROC curves of the best model (highlighted in Table 9) over all the models presented (single and ensemble) with the standard ESC 5-year SCD risk score applied to the same cohort (Figure 12). While our models showed slightly better discrimination, the improvement was modest, suggesting that although machine learning approaches hold promise, the added value over existing clinical tools remains limited in this first, external validation scenario.

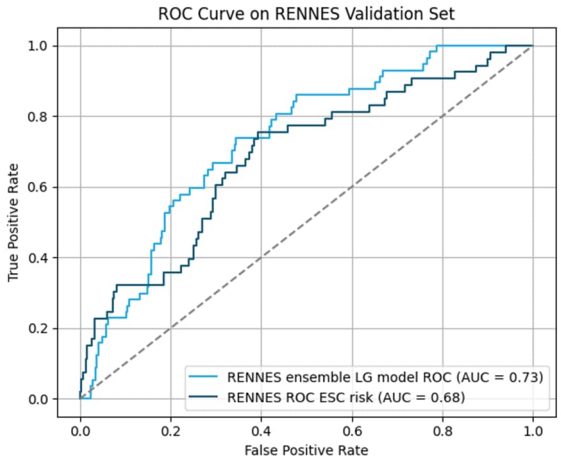

Figure 12: Ensemble model of 5 logistic regression model applied on the Rennes dataset of echocardiographic and clinical features (*in light blue*) compared to the ESC risk score (*in dark blue*).

In addition to ROC-based evaluation, we conducted a Kaplan Meier survival analysis on the Rennes cohort to assess whether the ensemble logistic regression model could meaningfully stratify patients based on predicted risk. Patients were grouped by the binary class output of the model, and survival curves were generated accordingly. The Kaplan Meier curves showed a clear separation between the two predicted groups, indicating that the model's predictions carry prognostic relevance over the follow-up period (Figure 13). A log-rank test confirmed a statistically significant difference in survival between the groups ($p = 1.756 \times 10^{-5}$). These results suggest that, despite modest classification performance, the model still captures relevant longitudinal risk information and may contribute to event stratification in clinical follow-up.

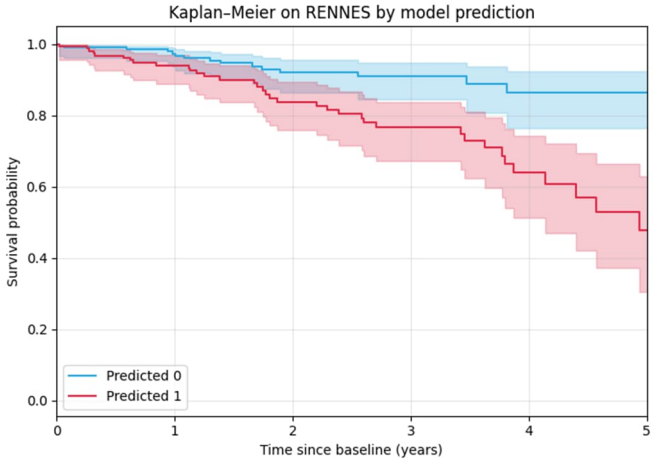

Figure 13: Kaplan Meier survival curves on the Rennes cohort stratified by the ensemble logistic regression model prediction (high: 1 vs. low predicted risk: 0).

2.4.2 Echocardiographic+Clinical+ECG Features

In a subsequent step, we extended the original echocardiographic model (subsection 2.3) by incorporating additional electrocardiographic (ECG) features routinely available in the EHR. Specifically, we added ECG-derived variables such as heart rate, rhythm, PR, QRS, and QT intervals. The dataset included 1,012 HCM patients from the SHARE registry with complete data across modalities. The same 5-year composite endpoint was used, encompassing heart failure progression and arrhythmic events.

The integration of these variables into the modelling pipeline led to improved performance across classifiers. As shown in Table 10, the random forest model achieved a balanced accuracy of 79.0% ± 1.5 and F1 score of 70.4% ± 2.1, while logistic regression reached balanced accuracy of 77.6% ± 1.2 and F1 score of 68.4% ± 1.3. These values represent an approximate 5% improvement in F1 score and a 7% gain in AUC compared to models using echocardiographic features alone.

Table 10: Cross-validated performance metrics (%) for the 2 best classifiers with the refined dataset of Florence and added all clinical features and ECG-derived features. Values are reported as mean ± standard deviation across the five outer folds.

Metric	Random Forest	Logistic Regression
Sensitivity	79.7 ± 4.1	79.4 ± 5.0
Specificity	78.3 ± 4.2	75.7 ± 3.9
Balanced Accuracy	79.0 ± 1.5	77.6 ± 1.2
F1 Score	70.4 ± 2.1	68.4 ± 1.3
AUC	87.0 ± 1.0	86.2 ±1.3

Receiver operating characteristic (ROC) curves for the updated models are presented in Figure 14 alongside the ESC risk score for comparison. While the ESC score remains a valuable reference, our models, demonstrated improved discrimination, with the added advantage of including several modalities.

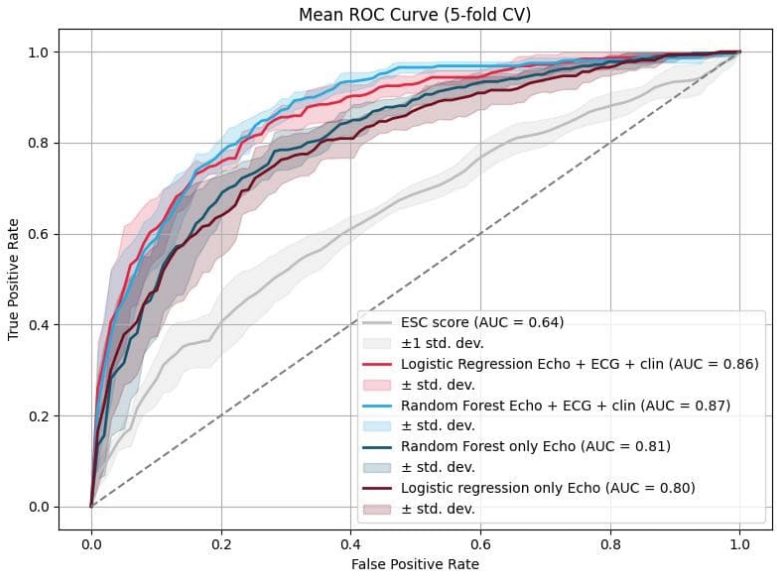

Figure 14: ROC curves of the 5-CV models with the addition of ECG-derived and clinical features, compare to the ESC risk score.

3 Summary of models and next steps

The following table summarizes the models developed up to the time of this Deliverable. For each model, it reports the dataset used for training (or fine-tuning) and the current stage of development (e.g., trained, cross-validated, externally validated). The column "DSS ready" indicates whether the model has already been implemented for integration into the Decision Support System (DSS) developed within WP7. The column "Pub" specifies whether the model has been the subject of a scientific publication in a journal or conference.

Table 11: Summary of developed data-driven models.

Model name	SMASH-HCM objective	Data required	Training Dataset	Advancement	DSS	Pub
ECG-Feature risk predictor	*Composite_VArrhyth.*	ECG-derived parameters	Rennes	Cross-validated on Rennes. External validation ongoing.	N	Y
ECG-Deep risk predictor	*Composite_VArrhyth.*	12 leads ECG	Rennes	Cross-validated on Rennes. External validation ongoing.	N	N
Echo-based	*Composite_Overall*	Echo-features	SHARE	Cross-validated on SHARE. External validation ongoing.	N	Y
Eco+clinical	*Composite_Overall*	Echo+Clinical features	SHARE	Cross-validated on SHARE. External validation ongoing.	N	N
Eco+Clinical+ECG	*Composite_Overall*	Echo+Clinical+ECG features	SHARE	Cross-validated on SHARE.	N	N
LVH classifier	Left Ventricular Hypertrophy Diagnosis	ECG-derived parameters	Publicly available	Ready	Y	Y
RHV classifier	Right Ventricular Hypertrophy Diagnosis	ECG-derived parameters	Publicly available	Ready	Y	Y
Multimodal cardiac	*Composite_HF, Composite_VArrhyth.*	ECG features + raw Cardiac MRI + Echo features	SHARE	Cross-validation in SHARE.	N	N
Multimodal cardiac + clinical features	*Composite_HF, Composite_VArrhyth.*	ECG features + raw Cardiac MRI + Echo features + Demographic + clinical visit history + genetic data + lab results + medications list	SHARE	Cross-validation in SHARE.	N	N
Multimodal cardiac + clinical features	*Composite_HF, Composite_VArrhyth.*	Raw ECG + Cardiac MRI + Echo features + Demographic + clinical visit history + genetic data + lab results + medications list	Rennes	Cross-validation in SHARE.	N	N
Multimodal cardiac and clinical features	*Composite_Overall, Composite_HF, Composite_VArrhyt.*	Sparse group subsets of: ECG features; cardiac MRI and echo features; and demographic, clinical, and lab variables	SHARE, [Rennes]	Cross-validated in SHARE and, where possible, externally validated in Rennes; production model fine-tuned in Rennes + SHARE	N	N

3.1 Models in progress

The table provides a snapshot of the status, while several ongoing activities aim to expand the pool of models. Current efforts focus on integrating additional modalities—such as cardiac MRI and ECG-derived biomarkers—to improve risk prediction, and on extending the models to multi-modal and multi-center settings. Data harmonization is actively being pursued to address inconsistencies in available features across cohorts: for example, the SHARE cohort includes MRI but not ECG, whereas the Rennes cohort provides ECG but lacks MRI. In addition, standardization challenges such as variable naming, measurement protocols, and handling of missing values are being progressively resolved, although some issues remain.

Future work will also explore ensemble and hybrid learning strategies to combine modality-specific predictions, including the use of simulated data or signals to augment limited modalities. This will enable richer and more personalized risk stratification tools for HCM care.

In parallel, new models will be developed to directly integrate heterogeneous modalities (e.g., raw cardiac MRI data alongside derived features), thereby reducing the need for modality-specific preprocessing. Examples of such approaches are outlined below.

Models using raw Cardiac MRI + features

This research aims to develop a Perceiver-based multimodal deep learning model for predicting survival probabilities in patients diagnosed with HCM at clinically relevant time points such as 1,3, and 5 years. The Perceiver architecture integrates heterogeneous multimodal data with minimal modality-specific processing, specifically 3D cardiac MRI volumes. It uses a cross-attention mechanism to efficiently reduce input dimensionality and latent Transformer to capture complex intermodal relationships for modelling high dimensional biomedical data. The first study focuses on three modalities (3D Cardiac MRI + Echo and ECG features) that directly reflect the cardiac structure and functions. The model will be trained to predict survival probabilities and composite clinical endpoints using only these three modalities. In a successive step, we include additional structured and unstructured clinical data. Each model will be evaluated individually and in combination to assess its predictive value. The best-performing subset will then be integrated with the core modalities to evaluate the benefits of extended multimodal fusion. Additional features from SHARE Dataset include Demographic, Clinical visit history, Genetic data (Text-based annotations), laboratory results, medications lists (Text-based annotations). To test the generalizability of the Perceiver-based multimodal model, we will conduct an external validation using Rennes dataset. The SHARE dataset includes ECG-derived features and raw CMR images, but Rennes dataset provides raw 12-lead ECG and CMR based features. For this reason, we need to adapt our model to accept raw ECG, Echo and CMR features with clinical features (similar to 4.2.4) for evaluation. This validation will help us to evaluate the performance across different populations and data sources.

3.2 Benchmarks models

Data-driven models from Tasks 6.3 and 6.4 will be benchmarked against literature-based models of HCM and its complications constructed in Task 6.1. The process used for constructing the initial library of such models is described in Deliverable 6.2. Models are automatically reconstructed from the literature using an enhanced instance of Pharmatics' biomedical literature-mining engine, MedAI, fine-tuned for HCM and related cardiovascular conditions. MedAI encodes biomedical texts with large language model (LLM) embeddings and applies cascade classifiers to filter sources by population, intervention, outcomes, and available variables aligned with our data dictionaries. From eligible sources, MedAI extracts and standardizes quantitative information, such as coefficients of statistical models, clinical decision rules, risk scores, and population characteristics tables, to construct a library

of *prior* predictive and risk-stratification models. These models serve as standalone baselines and can also provide priors, features, or architectural constraints for downstream data-driven learning.

During benchmarking, we plan to select MedAI-HCM prior models whose input variables align with the SMASH-HCM data dictionaries, such as those from the Rennes and SHARE datasets. Where suitable, we will recalibrate these models using individual-level data. To identify suitable models for benchmarking, we are developing an automated matching process. For a given target outcome (for example, sudden cardiac death) and a data dictionary, this process will rank candidate models based on how well their expected input variables match the variables in the SMASH-HCM dataset dictionaries. We will prioritize exact matches first, followed by variables that require only simple unit conversions, then credible proxy variables, and finally those requiring imputation. Models that cannot be feasibly mapped will be excluded from the benchmarking candidates. However, these models may still be used as sources of prior knowledge or weak supervision in Tasks 6.3 and 6.4.

When a selected function requires inputs that are missing in SMASH-HCM datasets, we will attempt to impute the missing values using principled approaches rather than simple population averages. For example, if a candidate benchmarking model uses five input variables and available individual-level data contains four with three overlapping, we will fill the gap using three complementary strategies. First, we will use equations or conditional models from the literature to predict missing values from observed variables. If this is not possible (for example, when published relationships are unavailable), we will estimate the missing variables using standard parametric or non-parametric regressors or classifiers trained across external datasets, using established ML pipelines. If the resulting performance metrics fall below predefined levels, we plan to apply Gaussian copulas. This involves first estimating marginal distributions for the missing variables using published population tables already extracted by Pharmatics' MedAI for HCM models; then estimating pairwise correlations for the latent variables corresponding to the missing observations from the literature or small datasets; and then conditioning on observed values to impute the missing ones in SMASH-HCM data in the corresponding Gaussian copula. Where relevant, we will use multiple imputation to express imputation uncertainty. We expect that the correlation estimation will need far less data than training full ML predictors for each missing variable.

As a set of benchmarking candidates, we will use models that achieve a minimum match score and show adequate calibration and prediction in internal validations on small data subsets after any required imputation. We will discard literature-based models that rely on extended missing variables or imputations not supported by data or literature. The identified benchmarking candidates will be compared with new data-driven models using appropriate evaluation metrics. These benchmarking models may also be included as predictors or experts in ensemble models, which will be considered in Task 6.4. All mappings, imputation methods, and assumptions will be documented, and calibrated functions will be made accessible via the API in WP7.

4 Conclusions

This document has presented the data-driven models developed so far, as well as those currently in progress, within the activities of WP6. Although the results remain preliminary—given that only partial data have been available—some of the developed models have already shown performance exceeding standard-of-care indices, such as the ESC score.

It is important to emphasize that these models are not intended to be used in isolation, but rather as part of a broader pool of models that also includes mechanistic approaches developed in WP3, WP4, and WP5. Their integration, which has just begun under Task 6.4, will be essential to fully exploit their complementary strengths in HCM patients' stratification, disease progression forecasting, and treatment response assessment.