Computers in Biology and Medicine 135 (2021) 104648

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A machine learning-based risk stratification model for ventricular tachycardia and heart failure in hypertrophic cardiomyopathy

Tim Smole [a], Bojan Žunkovič [a], Matej Pičulin [a], Enja Kokalj [a], Marko Robnik-Šikonja [a], Matjaž Kukar [a], Dimitrios I. Fotiadis [b], Vasileios C. Pezoulas [b], Nikolaos S. Tachos [b], Fausto Barlocco [c], Francesco Mazzarotto [c], Dejana Popović [d], Lars Maier [e], Lazar Velicki [f], Guy A. MacGowan [g], Iacopo Olivotto [c], Nenad Filipović [h], Djordje G. Jakovljević [g,i], Zoran Bosnić [a,*]

[a] University of Ljubljana, Faculty of Computer and Information Science, Večna Pot 113, Ljubljana, Slovenia
[b] University of Ioannina, Dept. of Materials Science and Engineering, Unit of Medical Technology and Intelligent Information Systems, Greece
[c] Cardiomyopathy Unit, Careggi University Hospital, University of Florence, Italy
[d] University of Belgrade, Clinic for Cardiology, Clinical Center of Serbia, Faculty of Pharmacy, Belgrade, Serbia
[e] University Hospital Regensburg, Dept. of Internal Medicine II (Cardiology, Pneumology, Intensive Care Medicine), Germany
[f] Faculty of Medicine, University of Novi Sad, Novi Sad, Serbia and Institute of Cardiovascular Diseases Vojvodina, Sremska Kamenica, Serbia
[g] Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle Upon Tyne, UK
[h] BIOIRC - Bioengineering Research and Development Center, Kragujevac, Serbia
[i] Faculty of Health and Life Sciences, Coventry University, Coventry, UK

ARTICLE INFO

Keywords:
Hypertrophic cardiomyopathy
Risk stratification
Machine learning
Artificial intelligence

ABSTRACT

Background: Machine learning (ML) and artificial intelligence are emerging as important components of precision medicine that enhance diagnosis and risk stratification. Risk stratification tools for hypertrophic cardiomyopathy (HCM) exist, but they are based on traditional statistical methods. The aim was to develop a novel machine learning risk stratification tool for the prediction of 5-year risk in HCM. The goal was to determine if its predictive accuracy is higher than the accuracy of the state-of-the-art tools.

Method: Data from a total of 2302 patients were used. The data were comprised of demographic characteristics, genetic data, clinical investigations, medications, and disease-related events. Four classification models were applied to model the risk level, and their decisions were explained using the SHAP (SHapley Additive exPlanations) method. Unwanted cardiac events were defined as sustained ventricular tachycardia occurrence (VT), heart failure (HF), ICD activation, sudden cardiac death (SCD), cardiac death, and all-cause death.

Results: The proposed machine learning approach outperformed the similar existing risk-stratification models for SCD, cardiac death, and all-cause death risk-stratification: it achieved higher AUC by 17%, 9%, and 1%, respectively. The boosted trees achieved the best performing AUC of 0.82. The resulting model most accurately predicts VT, HF, and ICD with AUCs of 0.90, 0.88, and 0.87, respectively.

Conclusions: The proposed risk-stratification model demonstrates high accuracy in predicting events in patients with hypertrophic cardiomyopathy. The use of a machine-learning risk stratification model may improve patient management, clinical practice, and outcomes in general.

1. Introduction

Hypertrophic cardiomyopathy (HCM) is the most prevalent disease among familial cardiomyopathies, and affects about one in 500 people. Most patients with HCM exhibit the “classic” hypercontractile HCM phenotype and have a stable course over the years, without evidence of heart failure (HF) progression. However, they remain at increased risk of life-threatening arrhythmias and sudden cardiac death (SCD) compared to the general population [1]. Pharmacological therapy fails to provide optimal protection, and high-risk patients generally receive an

* Corresponding author.
E-mail address: zoran.bosnic@fri.uni-lj.si (Z. Bosnić).

https://doi.org/10.1016/j.compbiomed.2021.104648
Received 13 May 2021; Received in revised form 8 July 2021; Accepted 8 July 2021
Available online 12 July 2021

implantable cardioverter-defibrillator (ICD) [2,3]. In some patients, HCM progresses towards progressive left ventricular (LV) dysfunction and refractory HF. Although such remodeling can take a long time (>10 years), it may be more precipitous and lead to death or cardiac transplantation even at a young age. Both SCD and HF progression are hard to predict based on presenting clinical and genetic features [4].

When assessing cardiomyopathy patients, risk stratification should be viewed as a dynamic, ongoing process based on the evaluation of patients' clinical and genetic features. With regard to SCD, high-risk status has been defined in several ways over the years, based on different international guidelines [5]. The most established risk factors for SCD in HCM include the family history of SCD, non-sustained ventricular tachycardia, abnormal blood pressure response to exercise, severe LV hypertrophy, and unexplained syncope [6]. Other variables are regarded as potential modifiers of this risk, including LV outflow tract obstruction (LVOTO) at rest, apical aneurysms, end-stage progression, extensive late gadolinium enhancement by magnetic resonance imaging, and complex genotypes [7–9]. Two leading risk stratification models currently exist. The model by O'Mahony et al. [10], incorporated in the 2014 European Society of Cardiology guidelines, is based on a retrospective analysis of a multicenter longitudinal cohort developed using the Cox proportional hazards model and validated using the bootstrapping method. The authors used a combination of eight factors (age, maximal LV wall thickness, left atrial (LA) diameter, LV outflow, family history of SCD, NSVT, and unexplained syncope) to predict patient-specific probabilities of SCD at five years. Their model and risk stratification tool is widely used and also available online. The alternative method, which is a part of the 2011 American College of Cardiology/American Heart Association guidelines [2,11] is based on consideration of individual risk factors, each associated with SCD in HCM at logistic regression analysis. Despite these advances, individualized prognostication remains challenging in HCM, with low specificity and positive predictive accuracy, independent of the score or algorithm used.

Machine learning (ML) has been recently proposed as a useful tool to improve management of disease prediction and progression [12]. Several ML approaches have been used in cardiology with the aim of improving the clinical workflow [13] and overcome the limitations of traditional methods. A recent example is a ML-based mortality prediction of patients undergoing cardiac resynchronization therapy (CRT) [14]. In this study, we develop and evaluate the first HCM risk stratification tool based on ML algorithms, considering patients' current clinical status, genetic data, imaging data, and medical history in order to identify patients at risk of any major adverse cardiac event (MACE) (including SCD and HF). To model the risk level (low-risk or high-risk), we apply four classification models (random forests, boosted trees, support vector machine, and neural networks) and evaluate their performance. Novelties and contributions of this paper include:

- description of the novel risk stratification system HCM-RSS that includes a proposal for the training data representation and performance evaluation of several supervised machine learning algorithms; and the performance of the system is compared to existing risk-stratification models for SCD, cardiac death and all-cause death risk-stratification,
- application of the SHAP methodology that explains the reasoning in the machine learning models that have the black-box nature and provides insight into influencing the parameters for risk stratification.

The paper is structured as follows: Section 2 provides the available data and presents its preprocessing, training with supervised machine learning, and explanation methodology. Section 3 presents the obtained results and compares the proposed approach with the state-of-the-art. Sections 4 and 5 conclude the paper.

2. Methods

2.1. Retrospective dataset with longitudinal information

The retrospective data that was used for training the machine learning algorithms was provided by the Careggi University Hospital, University of Florence, Italy. The used data were a part of a larger consortium, the Share Registry, and included longitudinal digital medical records from the entire hospital's medical practice. From all available data, data of 2302 patients (1448 male and 854 female) that were primarily diagnosed with HCM or had an HCM-diagnosed relative (inclusion criterion) were chosen for forming the machine learning dataset. The patients' records included their clinical data, management details, and disease-related events. Patients' demographic, physical, and clinical characteristics are given in Table 1. For each patient, the following data were collected: demographic and physical characteristics (gender, age, weight, and height); genetic data based on next generation sequencing-based testing; clinical investigations (echocardiography, electrocardiography, Holter monitoring, blood tests, cardiac magnetic resonance, and a cardiopulmonary exercise stress test); medication (medication type, date when the medication was started and stopped); and disease-related events (HF, VT, appropriate ICD intervention, SCD, cardiac death, and all-cause death). Since the dataset contains longitudinal clinical data from various clinical tests as well as relevant disease related events, it was possible to observe how patients' clinical status changed over time and label them as high- and low-risk patients.

2.2. Forming training data for machine learning

Due to practical reasons (i.e., the slow progression of the disease), a learning example is defined as a set of measurements that have been made within a time frame of one year. If the patient had a certain test performed multiple times within this time frame, the multiple tests are treated as separate measurements (see Fig. 1). If a certain type of test was not performed in the one-year time frame, the variables associated with that test were recorded as missing. Transforming patients' data in this way yielded a dataset with 13,386 learning examples (3.9 ± 4.8 examples per patient on average).

The dataset contains records of disease-related events (in total 4902 events) that occurred in patients with HCM along with the date of the occurrence (events are e.g., SCD, heart failure, and transplant). There are 97 different event categories present in the dataset. The main categories are: i) directly related to certain tests (e.g. abnormal Holter is a direct result of Holter monitoring test; pre-syncope may or may not be a result of stress test), ii) reflect a certain medical procedure (like pacemaker or ICD implantation, etc.); iii) represent patients' death (due to SCD, cardiac death, and all-cause death); and iv) represent what is referred to as quasi-death – if a life-saving treatment had not been provided, the patient would have died (e.g. heart transplant or ICD appropriate shock firing).

The following events were considered as high-risk events, hence the patients with such events were labeled as high-risk patients: abnormal Holter, ventricular tachycardia, non-sustained ventricular tachycardia (NSVT), sustained ventricular tachycardia (SVT), ventricular tachycardia ablation, abnormal exercise tolerance test (ETT), an implanted ventricular assist device (VAD), an implanted implantable cardioverter defibrillator (ICD), ICD appropriate firing, cardiac arrest, sudden cardiac death (SCD), heart failure (HF), the patient being listed for a heart transplant, and heart transplant. The risk label for each learning example was derived by verifying if any of the high-risk events has occurred to the patient in the 5-year time frame since the last measurement was taken. Some of the events are recorded only at the time a patient visits the clinic. Only those patients were included who had at least a 5-year follow-up, or, if the follow-up period was shorter, or they experienced at least one high-risk event.

Table 1
Patient demographic, physical and clinical characteristics. The upper part of the table shows patient overview characteristics and the lower part shows the statistics in terms of baseline and follow-up measurements.

Patients	Total (N = 2302)	
Characteristics	no. (%)	
Sex		
Male	1448 (62.9%)	
Female	854 (37.1%)	
Family history of HCM	983 (42.7%)	
Family history of SCD	426 (18.5%)	
Family history of CAD	104 (4.5%)	
Diabetes	82 (3.6%)	
Type 2 diabetes	73 (3.2%)	
Hypertension	214 (9.3%)	
Hypercholesterolemia	478 (20.8%)	
Genetic mutations	**Total tests performed (N = 1321)**	
MYBPC3	455 (34.4%)	
MYH7	254 (19.2%)	
MYL2	13 (1.0%)	
MYL3	7 (0.5%)	
TNNI3	42 (3.2%)	
TNNT2	45 (3.4%)	
TPM1	8 (0.6%)	
TTN	3 (0.2%)	
Measurements	**Baseline (N = 2302)**	**Follow-up (N = 1544)**
Characteristics	**no. (%)**	**no. (%)**
Age [years]	46 ± 19	54 ± 18
NYHA class		
I	1245 (54.1%)	750 (48.6%)
II	700 (30.4%)	552 (35.8%)
III	222 (9.6%)	208 (13.5%)
IV	11 (0.5%)	21 (1.4%)
Body mass index [kg/m^2]	25.2 ± 4.3	25.5 ± 3.9
Systolic blood pressure [mm Hg]	124.4 ± 19.4	122.2 ± 17.9
Diastolic blood pressure [mm Hg]	74.9 ± 10.2	74.4 ± 21.1
Left atrium [mm]	41.4 ± 8.8	44.2 ± 8.5
Left atrium volume [ml]	79.2 ± 39.6	93.2 ± 54.3
LVIDs	27.9 ± 6.7	29.2 ± 7.7
LVIDd	45.3 ± 6.7	46.3 ± 6.5
LVEF	65.1 ± 9.0	62.9 ± 9.5
Abnormal ETT	2 (0.1%)	27 (1.7%)
Abnormal holter	102 (4.4%)	351 (22.7%)
NSVT	106 (4.6%)	347 (22.5%)
SVT	20 (0.9%)	46 (3.0%)
Ventricular tachycardia	5 (0.2%)	14 (0.9%)
Ventricular tachycardia ablation	3 (0.1%)	5 (0.3%)
Atrial fibrillation	210 (9.1%)	350 (22.7%)
Atrial fibrillation ablation	22 (1.0%)	47 (3.0%)
Cardiac arrest	19 (0.8%)	29 (1.9%)
Heart failure	52 (2.3%)	93 (6.0%)
ICD implanted	76 (3.3%)	200 (13.0%)
ICD appropriate firing	9 (0.4%)	28 (1.8%)
Myectomy	28 (1.2%)	157 (10.2%)
Stroke	50 (2.2%)	78 (5.1%)
Pre-syncope	80 (3.5%)	123 (8.0%)
Syncope	184 (8.0%)	254 (16.5%)
Listed for heart transplant	0 (0.0%)	1 (0.1%)
Heart transplant	0 (0.0%)	2 (0.1%)

HCM - hypertrophic cardiomyopathy, SCD - sudden cardiac death, CAD - coronary artery disease, NYHA - New York Heart Association, LVIDS - left ventricular internal dimension at end-systole in mm/m2, LVIDd - left ventricular internal dimension at end-diastole in mm/m2, LVEF - left ventricular ejection fraction, ETT - exercise tolerance test, NSVT - non-sustained ventricular tachycardia, SVT - sustained ventricular tachycardia.

2.3. Data preprocessing

Datasets based on real world clinical practice, such as our dataset of 2302 patients, have a number of missing values and outliers, which represents an obstacle for ML. To facilitate the use of ML methods, we imputed the missing values [15,16] by applying the following procedures: (i) copying past/known values of the last result (in the range of five years) if the longitudinal data point is missing; (ii) adding values of healthy controls, i.e. the missing numerical values were replaced by random samples from the normal distributions, defined by the mean and standard deviation, which accurately describe the range of healthy values for a given numerical feature; (iii) resetting to mean values: for the numerical variables, where the standard (or healthy control) values are not available, the average value of each variable is used to impute its missing values. In the case of a categorical variable, the missing values are replaced with the most frequent value. Being aware that imputation of missing values could introduce data bias [17], we applied imputations (i) and (ii) as our first options, i.e., copying values from previous measurements of the same patient and using a variant of a multiple imputation method to keep the introduced data bias low. The highest data bias can thus be introduced with the imputation (iii), which we applied sparingly as our last option.

Furthermore, ML examples are often augmented by transformations that do not modify the label of the example, but change the input in a way that is realistic, which may improve the performance of the models. Generally, an example of such transformations can be, for example, translations and the rotations of images. For our dataset, we used the following augmentation techniques [18,19]: (i) Interpolation: many measurements were taken at non-equidistant time intervals and we used the time-based interpolation to compute approximations at regular periodic measurements. Parameter values and the target values between the two nearest measurements were linearly interpolated. Categorical (ordinal) values, which are represented as integers, were rounded after the interpolation; (ii) Adding longitudinal data: we used the available patients' longitudinal data to extend the feature space by combining all possible pairs of patient measurements. The risk label of such a longitudinal patient record is determined by the label of the latest measurements. This procedure significantly increases the number of training examples (from initial 3465 to approx. 10^5); and (iii) Semi-supervised learning exploits the fact that we are in possession of many more unlabeled instances than the labeled ones. Using the available labeled examples, we built the prediction model that predicted labels for the unlabeled example. The most reliable of such examples were added to the training set and helped to improve the predictive model that was rebuilt using this expanded dataset.

For training and evaluation of ML models, all available examples were used (i.e., none were discarded due to too many missing values). However, for use of HCM-RSS in production environment, some basic parameters were selected as mandatory to be provided for learning and risk stratification to run (such as gender, age, dates of disease-related events etc.).

2.4. Predictive modeling with supervised machine learning

To model risk-level of patients, we applied four different ML models for supervised learning: (i) Random forests [20,21], which are an ensemble prediction model that constructs multiple decision trees (implementation in the statistical package R and Python Scikit-Learn [22]); (ii) XGBoost library [23], which provides gradient boosting on an ensemble of many decision trees; (iii) Support Vector Machines [24] classifier, which uses linear or radial basis kernel functions to separate the two classes (implementation in R package *e1071* was used); and (iv) fully-connected feed-forward neural networks, which mimic the working of neurons in brain. The best parameter configuration (hyper parameters) for all used classification models were optimized using the Bayesian optimization and random search implemented in *keras-tuner*

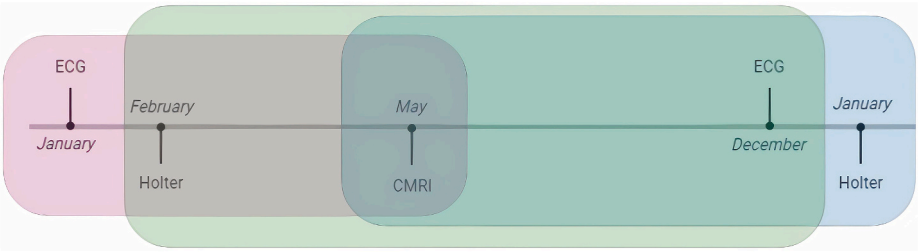

Fig. 1. Construction of a learning example: The graphics shows three learning examples (denoted by pink, green, and blue rounded rectangles) constructed from five clinical tests (measurements) all performed within 13 months: ECG (January), Holter (February), CMRI (May), ECG (December), and Holter (January next year – not within one year of the first measurement). The clinical tests are combined into examples within the timeframes of one year.

[25]. The final selected parameters were as follows:

- Random forest: 1000 trees, entropy as attribute selection criterion, maximum depth of a tree 75, the number of considered features equal to log2 of the number of available features, bootstrapping not used;
- XGBoost: 1000 gradient-boosted trees, *gbtree* booster, learning_rate 0.2, maximum depth 4, subsample ratio of training instances 0.75, gamma (minimum loss reduction for partitioning a leaf) 1.5,
- Support Vector Machines: linear kernel, gamma = 1/(number_of_attributes), epsilon = 0.1, C = 1,
- Neural-networks: learning rate 0.002, dropout probability 0.3, regularization strength 0.001, 5 hidden layers, *elu* activation functions in hidden layers, sigmoid activation functions in the output layer, 200-200-200-20-20 neurons in hidden layers.

For the remaining parameter values (if any) default values were used.

2.5. Explanation of the predictive model

Most of the top performing machine learning models are black boxes, which means that the prediction-making process is not transparent to the user. In risk-sensitive areas such as healthcare, model interpretability is of a crucial importance, because users are not only interested in good predictive accuracy, but also in understanding the decision process – which is what makes the predictions actionable. Explanation methods provide information about why a certain prediction was made for a patient and which features had the highest impact in making the prediction [26].

The risk stratification model classifies patients into high-risk or low-risk classes, and for each prediction an explanation is generated. An explanation consists of a list of the most relevant features that influence the prediction, either supporting the predicted class or opposing it. This type of explanation for individual patients eases understanding of the reasoning process captured by the ML model. Summarizing these explanations over the whole dataset provides further insights into the model's behavior and allows for qualitative understanding of the relationship between the patient's features and the model's prediction. We used a model-agnostic explanation method/library SHAP [27] (SHapley Additive exPlanations) that is based on the theoretically well-supported concept of Shapley values from cooperative game theory. Since our best performing model is a boosted tree, we used a variant of SHAP adapted to these models.

2.6. Statistical methods

We performed the model evaluation by computing the classification accuracy (ratio of correctly classified examples), sensitivity (true positive rate), specificity (true negative rate), precision (positive predictive value), F_1 score (harmonic mean between precision and sensitivity), and AUC for the learned predictive models. We computed all metrics using the stratified 10-fold cross-validation procedure, which uses multiple data samples to unbiasedly estimate how the model will perform with independent data. The ROC (Receiver Operating Characteristic) curve is a two-dimensional representation of a classifier performance, which depicts the true positive rate given the false positive rate. The area under the ROC curve (AUC) gives the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example [28].

3. Results

3.1. Performance of predictive models

The testing performance of the used classification models is shown in Table 2. The best results were achieved using the boosted trees, which obtained an average accuracy of 0.75, AUC 0.82, and an F_1 score of 0.71.

We further analyzed the performance of boosted trees for each high-risk event separately, where the positive class contained only one high-risk event, the examples with other high-risk events were excluded, and the negative class remained the same. Fig. 2 shows the ROC curves that summarize the performance of the risk stratification model for specific high-risk events. For reference, the results for prediction of all events, performed by the original model and denoted with "All events," are included, as well. Based on the results, we can group the events into three categories:

1. $\boldsymbol{AUC} > 0.85$: VT, HF, ICD appropriate firing,
2. $0.76 < \boldsymbol{AUC} < 0.85$: abnormal ETT, abnormal holter, NSVT, ICD appropriate firing, high risk (all events),
3. $\boldsymbol{AUC} < 0.76$: SCD.

For the first group of events, our model performs better than the joint ("All events") high-risk classifier. The most difficult task is the prediction of SCD events in patients without ICD implants. This was expected due to the stochastic nature of the SCD events. Several high-risk events (*cardiac arrest*, *listed for transplant*, *transplant*, *VAD*, and *VT ablation*) could not be evaluated with this procedure due to the low number of instances in the dataset.

3.2. Comparison with the state-of-the-art

Due to the lack of HCM risk stratification approaches, we compare the developed model with the most similar available risk stratification models that are relevant to HCM patients: (i) use of conventional risk factors in clinical practice [2,3], (ii) prediction of SCD [10], (iii) prediction of cardiac death (CD), and (iv) prediction of all cause death (AD). In the case of the SCD, we compare the performance of our model with the state-of-the-art SCD calculator. For the cardiac death and any death, we construct risk-factor models based on clinical guidelines. The comparisons are described below and summarized in Table 3. Additionally, ROC curves for all compared models are shown in Fig. 3. Besides the improvement with respect to the manual high-risk identification model

Table 2
Performance of the machine learning algorithms on the task of risk stratification of HCM patients. The results of the 10-fold cross-validation for predicting high-risk patients five years ahead are shown. The reported values are mean values and standard deviation between cross-validation folds. The best results for each metric are in bold.

Model	Accuracy	AUC	Specificity	Sensitivity	Precision	F_1 score
Random forest	0.72 ± 0.03	0.79 ± 0.03	0.81 ± 0.05	0.62 ± 0.03	0.74 ± 0.05	0.68 ± 0.03
SVM (linear)	0.69 ± 0.05	0.74 ± 0.04	0.69 ± 0.05	**0.69 ± 0.08**	0.59 ± 0.08	0.63 ± 0.07
SVM (RBF)	0.67 ± 0.02	0.73 ± 0.03	0.68 ± 0.03	0.64 ± 0.05	0.62 ± 0.04	0.63 ± 0.04
Boosted trees	**0.75 ± 0.02**	**0.82 ± 0.02**	0.81 ± 0.03	0.67 ± 0.04	0.78 ± 0.02	**0.72 ± 0.02**
Neural-Networks	0.74 ± 0.03	0.80 ± 0.04	**0.86 ± 0.05**	0.61 ± 0.07	**0.79 ± 0.05**	0.68 ± 0.05

AUC – Area Under Curve, SVM – support vector machine, RBF – radial basis kernel.

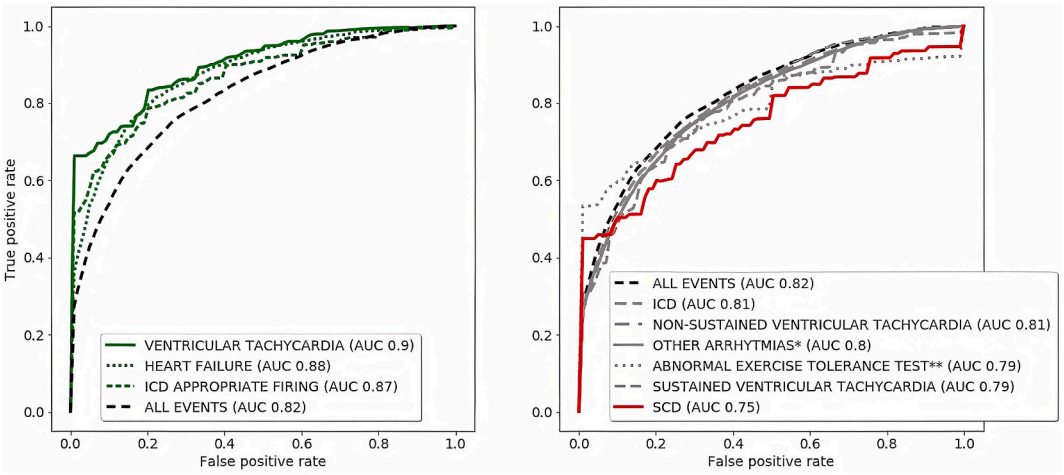

Fig. 2. ROC curves for classification of specific high-risk events. The left and right panels displays ROC curves for high-risk events that perform better and worse than the model for all events, respectively.
* atrial fibrillation, non-sustained ventricular tachycardia, 2nd or 3rd degree AV-block
** defined as failure to increase blood pressure (<30 mmHg increase) or presence of fall in blood pressure during effort.

Table 3
A comparison of the HCM-RSS with the specific risk-stratification models. Means and standard deviations of metrics are reported. The best results are denoted in bold.

Model	Accuracy	AUC	Specificity	Sensitivity	Precision	F_1
CONV	0.60 ± 0.001	0.60 ± 0.001	0.57 ± 0.001	0.65 ± 0.01	0.60 ± 0.002	0.62 ± 0.001
HCM-RSS (CONV)	**0.75 ± 0.02**	**0.82 ± 0.02**	**0.81 ± 0.03**	**0.67 ± 0.04**	**0.78 ± 0.02**	**0.72 ± 0.02**
SCDcalc1	0.58 ± 0.03	0.58 ± 0.03	0.72 ± 0.05	0.44 ± 0.04	0.62 ± 0.04	0.51 ± 0.01
SCDcalc2	0.60 ± 0.02	0.60 ± 0.02	**0.87 ± 0.04**	0.32 ± 0.03	**0.72 ± 0.06**	0.45 ± 0.01
HCM-RSS (SCD)	**0.66 ± 0.08**	**0.70 ± 0.1**	0.64 ± 0.13	**0.69 ± 0.16**	0.66 ± 0.07	**0.66 ± 0.1**
CD	**0.64 ± 0.01**	0.64 ± 0.01	**0.87 ± 0.01**	0.4±1e-10	**0.77 ± 0.02**	**0.65 ± 0.01**
HCM-RSS (CD)	0.60 ± 0.07	**0.70 ± 0.12**	0.52 ± 0.15	**0.69 ± 0.15**	0.59 ± 0.06	0.63 ± 0.08
AD	**0.66 ± 0.01**	0.70 ± 0.01	**0.68 ± 0.02**	0.64 ± 0.01	**0.67 ± 0.02**	**0.65 ± 0.01**
HCM-RSS (AD)	0.65 ± 0.09	**0.71 ± 0.12**	0.64 ± 0.13	**0.67 ± 0.15**	0.65 ± 0.1	**0.65 ± 0.1**

CONV – conventional risk factors, SCD – SCD calculator, CD – cardiac death, AD – all cause death.

(HR), the results show improvements in all specific tasks (SCD, CD, and AD).

3.2.1. Comparison with conventional risk factors

In clinical practice, the risk score can be constructed by counting how many conventional risk factors of SCD, HF, and other high risk events are present [2,3]. To apply such an approach on our data, we alternatively predicted the risk level by counting such conventional risk factors that are available in the given dataset (we denote this approach with CONV). Here, the following risk factors were considered as high-risk [29]: NYHA class $\geq$ III, NSVT (defined as $\geq$ 3 consecutive ventricular beats at a rate of >120 beats per minute), interventricular (IVSs) or posterior systolic wall thickness (PWTs) > 30 mm, restrictive LV filling pattern (diastolic dysfunction grade 3), family history of SCD, unexplained syncope, LA diameter $\geq$ 48 mm, LV ejection fraction (LVEF) $\leq$ 50%, LVOT peak gradient at rest > 30 mmHg, N-terminal pro-B-type natriuretic peptide (NT-proBNP) > 900 pg/ml, and atrial fibrillation in any form (AF).

The experiments revealed that the optimal CONV approach on our dataset yields the higher accuracy if the threshold of two present risk factors is used to classify the patient as a high-risk patient. We used this best-performing CONV model to compare its accuracy to the accuracy of our model HCM-RSS, the comparison is summarized in Table 3. The results show that we can observe an approximately 37% AUC improvement of HCM-RSS compared with the model CONV, both models having AUCs of 0.82 and 0.60, respectively.

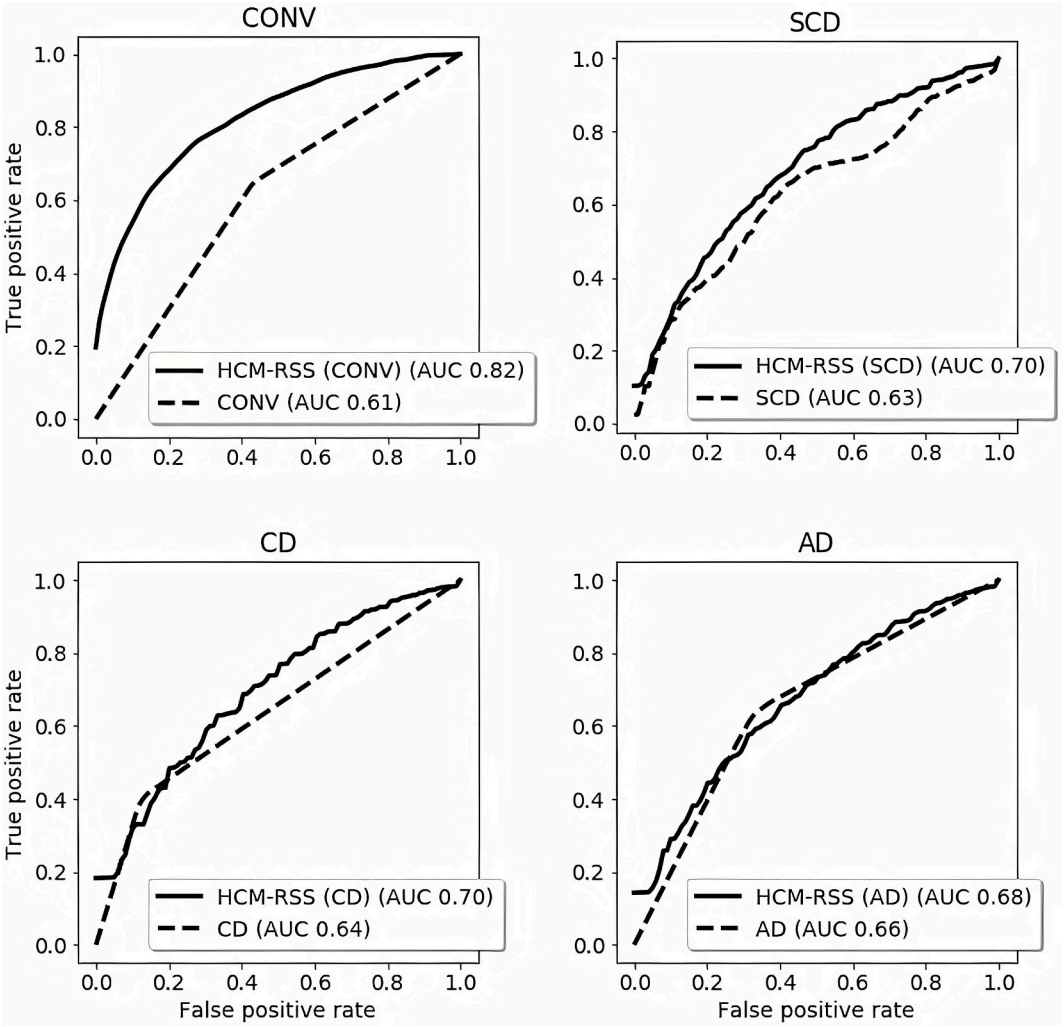

Fig. 3. Comparison of ROC curves for HCM-RSS (restricted to predicting only a given task) and other existing approaches: model that uses conventional risk factors (CONV), SCD calculator (SCD), cardiac death model (CD), and the all cause death (AD) model.

3.2.2. *Comparison with the SCD calculator*

The purpose of the SCD calculator [10] is to estimate the risk of SCD in five years for HCM patients, and it returns three risk categories: ICD not indicated, ICD maybe indicated, or ICD indicated. The score is calculated as $1 - 0.998 \cdot exp(\textit{Prognostic Index})$, where

$$\begin{aligned} \textit{Prognostic Index} = {} & 0.15939858 \cdot \textit{max LV wall thickness}(mm) \\ & -0.00294271 \cdot \textit{max LV wall thickness}^2(mm^2) \\ & +0.0259082 \cdot \textit{left atrial diameter}(mm) \\ & +0.00446131 \cdot \textit{max LVOT gradient}(mmHg) \\ & +0.4583082 \cdot \textit{Family history SCD} \\ & +0.82639195 \cdot \textit{NVST} \\ & +0.71650361 \cdot \textit{Unexplained syncope} \\ & +0.01799934 \cdot \textit{Age (years)}. \end{aligned}$$

Since SCD is one of the considered high-risk events, we were able to compare the performance of the SCD calculator with the performance of the HCM-RSS model only by predicting the SCD. To do that, we applied the SCD calculator to our database and compared the results with the output of our model. Two strategies for assigning a high-risk class to the patient were used. In the first (denoted as 'SCDcalc1' in Table 3), a high-risk class was assigned to patients labeled on the calculator as either '*ICD maybe indicated*' or '*ICD indicated.*' In the second strategy (denoted as 'SCDcalc2' in Table 3), a high-risk class was only assigned to patients labeled with the calculator as '*ICD indicated.*' The results in Table 3 show that only for the SCD task the restricted HCM-RSS model outperforms both variants of the SCD calculator (SCDcalc1 and SCDcalc2), all three models having AUCs of 0.70, 0.58, and 0.60, respectively.

3.2.3. *Comparison with the cardiac death and all cause death risk-stratification models*

Similarly as for the SCD, we can compare our model with the models obtained based on the risk factors for cardiac death and all cause death (denoted as CD and AD, respectively). Based on the clinical guidelines [29], we consider the following risk factors: the NYHA functional class, the family history of SCD, syncope, AF, NSVT, maximal LV wall thickness, and obstruction (LVOT peak gradient at rest > 30 mmHg). The best performing alternative CD model was obtained by considering at least two risk factors, and the best alternative AD model was obtained by considering at least three or more risk factors. We calculated the performance metrics on our dataset in both cases and compared them with HCM-RSS that was restricted to predicting only CD or AD outcomes. The results show that our model performs on par with the risk-factor baseline models, despite being trained on a very different task (see Table 3).

3.3. *Explanation of the risk stratification model*

The visualization of a model explanation consists of a horizontal bar chart that is divided into multiple subsections, each pertaining to a specific feature. The subsections contain horizontal bars that can extend either to the left (negative, red), to the right (positive, green), or both. The latter is due to averaging across many learning examples that are

situated in different contexts, and so the effect and significance of feature values vary. The length of the bars corresponds to the magnitude of the impact, i.e. the longer the bar the bigger the impact; the color corresponds to the direction of the impact, i.e. green for positive (supporting the prediction of the target class), and red for negative (opposing it). Each subsection includes a hatched top bar that represents the average contribution of a feature, and below it there are the contributions grouped by feature values, i.e. Yes/No for binary features, and three discretization intervals for numeric features.

The visualization of the model explanation for the predicted class "high-risk" is shown in Fig. 4. It was obtained by averaging the features' contributions over all test set instances across all 10 folds. The horizontal bars denote the features' impact on the prediction, and they are sorted by the absolute sum value of the overall contribution of each feature (positive and negative). In this way, the graph allows the user to compare the direction of the impact (positive/negative), as well as its magnitude for individual feature values. We only visualized the features for which the sum of contributions encompasses at least 50% of all contributions (filter out of less impactful features). The explanation shows that the prediction of the high-risk class is in greatest magnitude (the features with the largest impact are highlighted):

- supported by features values such as:
  1. presence of tricuspid regurgitation,
  2. presence of arrhythmogenic right ventricular cardiomyopathy (ARVC),
  3. LA dimension >48.00 mm,
  4. LA volume >89.00 ml,

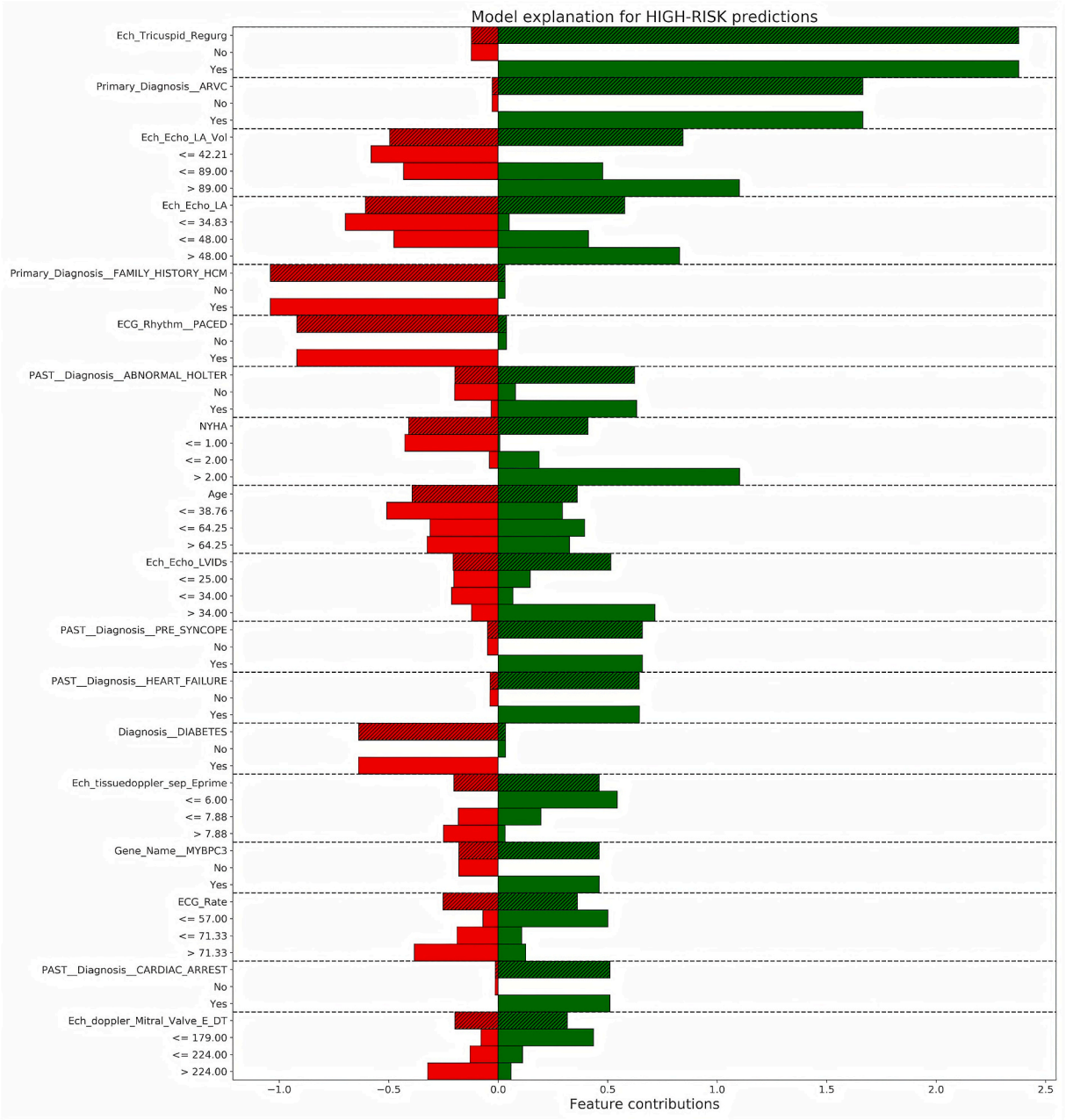

Fig. 4. Model explanation for identification of high-risk patients. The feature values that increased the prediction probability of the high-risk class are shown in green and those that decreased it are shown in red. On the x-axis, the magnitude of the feature value's impact is shown and corresponds to the length of the bars, i.e. the longer the bar the bigger the impact. The y-axis represents the discretized features and their values that significantly contributed to the prediction of the high-risk class. Each feature's average (positive and negative) impact is shown with hatched bars.

5. NYHA class >2;

- opposed by features values such as:
  1. Age ≤38.76 years,
  2. LA diameter ≤34.83 mm,
  3. LA volume ≤42.21 ml,
  4. family history for HCM,
  5. pacemaker presence,
  6. diabetes presence.

4. Discussion

The aim of the study is to explore machine learning methods as a potential tool for extracting knowledge from the patients' data, and model the level of HCM risk. Although machine learning has already been used in cardiology for mortality prediction, only the traditional statistical methods prevail for predicting the risk of HCM. To fill this gap, we developed a novel HCM risk stratification model using machine learning methods, called 'HCM-RSS.' Our developed model considers patients' current clinical status, genetic data, imaging data, and medical history in order to identify patients at risk for any major adverse cardiac event - MACE (including SCD and HF). To model the risk level (low-risk or high-risk), we apply four classification models (random forests, boosted trees, support vector machine, and neural networks), and evaluate their performance.

The major findings suggest that the boosted trees model demonstrates high predictive accuracy (accuracy of 75%, AUC 0.82, and F_1 score 0.71). Detailed analysis revealed that it performs best (with AUC > 0.85) for predicting ventricular tachycardia, heart failure, and ICD triggering/activation. Our further comparison of the HCM-RSS's performance with related risk predictors [29] for HCM (using conventional risk factors), SCD [10], cardiac death, and all cause death, revealed the favorable performance advantage of the HCM-RSS. In particular, the HCM-RSS achieved higher AUC than using conventional risk factory by 37%, SCD calculator by 17%, the cardiac death model by 9%, and the all cause death model by 1%.

O'Mahony et al. [10]' were among the first who proposed a validated risk prediction model for SCD in HCM patients. They emphasized that the current clinical guidelines for HCM are based on the summation of a limited number of binary parameters, denoting if a single type of risk is present or not. This can represent a problem, since binary variables can provide a more crude approximation of the computed risk compared to models that can consider continuous variables. Use of machine learning classifiers for HCM-RSS allowed us to overcome this limitation, presumably yielding higher risk predictive accuracy, including for SCD. Additionally, Steriotis and Sharma (2015) [6] reported that the major SCD risk factors (given 2003 ACC/ESC and 2011 ACCF/AHA Guidelines), which include unexplained syncope, family history of SCD, severe left ventricular hypertrophy, non-sustained VT, and an abnormal blood pressure response to exercise have low positive predictive accuracy, as the disease is clinically very diverse. As the machine learning predictive modeling was able to model non-linear dependencies between patient parameters, this study opens doors for determining more complex interplay between parameter values when predicting risk.

When surveying the use of machine learning algorithms (neural networks, support vector machines, random forests) in cardiology, Cuocolo et al. [13]' reported that their accuracy on imaging problems ranges from 87.2%–97.8% (AUC ranging from 0.73 to 0.95). Although HCM risk stratification is a different domain, it is interesting to see that the HCM-RSS reached a comparable AUC of 0.82 of numerical data, of which some are also derived from imaging.

Tsay and Patterson (2018) [12] have highlighted that major limitations of machine learning algorithms in the medicine stem from the inflated expectations about their ease of implementation. They indicated that one the most problematic aspects is the black-box nature of machine learning models, which obstructs insight into automatically proposed decisions and reasons for the proposed interventions. Since the transparency, justifiability, and understanding of the decision process are important in risk-sensitive areas such as cardiology, we aimed at overcoming this drawback by applying the SHAP-based explanation method. The automatically generated visualization extracted and revealed several HCM risk factors, which play a decisive role in the models' risk prediction. Some of these factors also confirmed what is already known about the HCM, such as positive predictive influences of the parameters: the presence of tricuspid regurgitation, the presence of arrhythmogenic right ventricular cardiomyopathy (ARVC), LA dimension > 48.00 mm, LA volume > 89.00 ml, and NYHA class > 2.

4.1. Limitations and further work

The development of HCM-RSS addressed several machine learning challenges [16,18,19] by utilizing a sequence of data preprocessing methods (imputing, matching longitudinal records, using unlabeled examples, and the transformation to training examples for machine learning). Generally, such preprocessing steps are tailored to the data that are available at hand, and their choice can determine the resulting accuracy of the predictive models. Nevertheless, it would make sense to determine the guidelines, how different data transformations work for different machine learning problems in cardiology. Likewise, a sensitivity study of the results shall be performed to determine how the patient's record timeframe and predicted risk timeframe influence the achieved accuracies.

Secondly, the study assumed that all data are available prior to utilizing machine learning and that the trained model would not change over time. It would also make sense to evaluate the accuracy of different methods for incremental updates of the model with the new patients' data to determine if and how the risk factors change over time.

Importantly, although the models were cross-validated on the data from the same hospital, the further work shall include validation on external data from other hospitals or centers. The challenges with this aim lie in the data record format that is not unified across all medical institutions, and the availability of the same parameters that were used to design the approach in this paper.

To conclude, the HCM-RSS represents a promising approach to improve risk stratification accuracy in hypertrophic cardiomyopathy. In future work, we plan to extend this work to developing predictive models for disease progression. We hope that the developed approaches will provide automated diagnostic reference and early identification of high-risk patients, which will open new opportunities for knowledge acquisition in clinical cardiology using artificial intelligence approaches.

Funding

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 777204 (www.silicofcm.eu). This article only reflects the author's view. The Commission is not responsible for any use that may be made of the information it contains.

Data statement

The data underlying this article cannot be shared publicly to ensure the privacy of patients from the Careggi University Hospital, Florence, Italy. The data will be shared on reasonable request to the corresponding author. The final obtained decision model will also be shared on reasonable request to the corresponding author.

Declaration of competing interest

None declared.

References

[1] I. Olivotto, F. Cecchi, C. Poggesi, M.H. Yacoub, Patterns of disease progression in hypertrophic cardiomyopathy: an individualized approach to clinical staging, Circ. Heart Fail. 5 (2012) 535–546, https://doi.org/10.1161/CIRCHEARTFAILURE.112.967026.

[2] B.J. Gersh, B.J. Maron, R.O. Bonow, J.A. Dearani, M.A. Fifer, M.S. Link, S.S. Naidu, R.A. Nishimura, S.R. Ommen, H. Rakowski, C.E. Seidman, J.A. Towbin, J. E. Udelson, C.W. Yancy, 2011 ACCF/AHA guideline for the diagnosis and treatment of hypertrophic cardiomyopathy: executive summary: a report of the American College of cardiology foundation/American heart association task force on practice guidelines, Circulation 124 (2011) 2761–2796, https://doi.org/10.1161/CIR.0b013e318223e230.

[3] B.J. Maron, W.J. McKenna, G.K. Danielson, L.J. Kappenberger, H.J. Kuhn, C. E. Seidman, P.M. Shah, W.H. Spencer, P. Spirito, F.J. Ten Cate, E.D. Wigle, R. A. Vogel, J. Abrams, E.R. Bates, B.R. Brodie, P.G. Danias, G. Gregoratos, M. A. Hlatky, J.S. Hochman, S. Kaul, R.C. Lichtenberg, J.R. Lindner, R.A. O'Rourke, G. M. Pohost, R.S. Schofield, C.M. Tracy, W.L. Winters, W.W. Klein, S.G. Priori, A. Alonso-Garcia, C. Blomström-Lundqvist, G. De Backer, J. Deckers, M. Flather, J. Hradec, A. Oto, A. Parkhomenko, S. Silber, A. Torbicki, American College of cardiology/European society of cardiology clinical expert consensus document on hypertrophic cardiomyopathy: a report of the American College of cardiology foundation task force on clinical expert consensus documents and the European society of cardiology committee for practice guidelines, in: J. Am. Coll. Cardiol, Elsevier Inc., 2003, pp. 1687–1713.

[4] G. Makavos, C. Kairis, M.E. Tselegkidi, T. Karamitsos, A.G. Rigopoulos, M. Noutsias, I. Ikonomidis, Hypertrophic cardiomyopathy: an updated review on diagnosis, prognosis, and treatment, Heart Fail. Rev. 24 (2019) 439–459, https://doi.org/10.1007/s10741-019-09775-4.

[5] I. Christiaans, K. Van Engelen, I.M. Van Langen, E. Birnie, G.J. Bonsel, P.M. Elliott, A.A.M. Wilde, Risk stratification for sudden cardiac death in hypertrophic cardiomyopathy: systematic review of clinical risk markers, Europace 12 (2010) 313–321, https://doi.org/10.1093/europace/eup431.

[6] A.K. Steriotis, S. Sharma, Risk stratification in hypertrophic cardiomyopathy, Eur. Cardiol. 10 (2015) 31–36, https://doi.org/10.15420/ecr.2015.10.01.31.

[7] B.J. Maron, E.J. Rowin, S.A. Casey, T.S. Haas, R.H.M. Chan, J.E. Udelson, R. F. Garberich, J.R. Lesser, E. Appelbaum, W.J. Manning, M.S. Maron, Risk stratification and outcome of patients with hypertrophic cardiomyopathy ≥60 years of age, Circulation 127 (2013) 585–593, https://doi.org/10.1161/CIRCULATIONAHA.112.136085.

[8] E.J. Rowin, M.S. Maron, The role of cardiac MRI in the diagnosis and risk stratification of hypertrophic cardiomyopathy, Arrhythmia Electrophysiol. Rev. 5 (2016) 197–202, https://doi.org/10.15420/aer.2016:13:3.

[9] E.T.D. Hoey, J.K. Teoh, I. Das, A. Ganeshan, H. Simpson, R.W. Watkin, The emerging role of cardiovascular MRI for risk stratification in hypertrophic cardiomyopathy, Clin. Radiol. 69 (2014) 221–230, https://doi.org/10.1016/j.crad.2013.11.012.

[10] C. O'Mahony, F. Jichi, M. Pavlou, L. Monserrat, A. Anastasakis, C. Rapezzi, E. Biagini, J.R. Gimeno, G. Limongelli, W.J. McKenna, R.Z. Omar, P.M. Elliott, A novel clinical risk prediction model for sudden cardiac death in hypertrophic cardiomyopathy (HCM Risk-SCD), Eur. Heart J. 35 (2014) 2010–2020, https://doi.org/10.1093/eurheartj/eht439.

[11] M.S. Maron, E.J. Rowin, B.S. Wessler, P.J. Mooney, A. Fatima, P. Patel, B. C. Koethe, M. Romashko, M.S. Link, B.J. Maron, Enhanced American College of cardiology/American heart association strategy for prevention of sudden cardiac death in high-risk patients with hypertrophic cardiomyopathy, JAMA Cardiol. 4 (2019) 644–657, https://doi.org/10.1001/jamacardio.2019.1391.

[12] D. Tsay, C. Patterson, From machine learning to artificial intelligence applications in cardiac care: real-world examples in improving imaging and patient Access, Circulation 138 (2018) 2569–2575, https://doi.org/10.1161/CIRCULATIONAHA.118.031734.

[13] R. Cuocolo, T. Perillo, E. De Rosa, L. Ugga, M. Petretta, Current applications of big data and machine learning in cardiology, J. Geriatr. Cardiol. 16 (2019) 601–607.

[14] M. Tokodi, W.R. Schwertner, A. Kovács, Z. Tősér, L. Staub, A. Sárkány, B. K. Lakatos, A. Behon, A.M. Boros, P. Perge, V. Kutyifa, G. Széplaki, L. Gellér, B. Merkely, A. Kosztin, Machine learning-based mortality prediction of patients undergoing cardiac resynchronization therapy: the SEMMELWEIS-CRT score, Eur. Heart J. 41 (2020) 1747–1756, https://doi.org/10.1093/eurheartj/ehz902.

[15] Y. Liu, V. Gopalakrishnan, An overview and evaluation of recent machine learning imputation methods using cardiac imaging data, Data 2 (2017) 8, https://doi.org/10.3390/data2010008.

[16] A.K. Waljee, A. Mukherjee, A.G. Singal, Y. Zhang, J. Warren, U. Balis, J. Marrero, J. Zhu, P.D.R. Higgins, Comparison of imputation methods for missing laboratory data in medicine, BMJ Open 3 (2013), e002847, https://doi.org/10.1136/bmjopen-2013-002847.

[17] T.R. Sullivan, A.B. Salter, P. Ryan, K.J. Lee, Bias and precision of the "multiple imputation, then deletion" method for dealing with missing outcome data, Am. J. Epidemiol. 182 (2015) 528–534, https://doi.org/10.1093/aje/kwv100.

[18] M.B. Landrum, M.P. Becker, A multiple imputation strategy for incomplete longitudinal data, Stat. Med. 20 (2001) 2741–2760, https://doi.org/10.1002/sim.740.

[19] C. Genolini, A. Lacombe, R. Écochard, F. Subtil, CopyMean: a new method to predict monotone missing values in longitudinal studies, Comput. Methods Progr. Biomed. 132 (2016) 29–44, https://doi.org/10.1016/j.cmpb.2016.04.010.

[20] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.

[21] J.D. Malley, J. Kruppa, A. Dasgupta, K.G. Malley, A. Ziegler, Probability Machines: consistent probability estimation using nonparametric learning machines, Methods Inf. Med. 51 (2012) 74–81, https://doi.org/10.3414/ME00-01-0052.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M.P.É. Duchesnay, Scikit-learn: machine learning in Python. http://scikit-learn.sourceforge.net, 2011. (Accessed 1 January 2021).

[23] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min, Association for Computing Machinery, New York, NY, USA, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[24] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297, https://doi.org/10.1007/bf00994018.

[25] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi, Others, Keras Tuner, 2019. https://github.com/keras-team/keras-tuner.

[26] M. Robnik-Šikonja, M. Bohanec, Perturbation-based explanations of prediction models, in: Hum. Mach. Learn, Springer, Cham, 2018, pp. 159–175, https://doi.org/10.1007/978-3-319-90403-0_9.

[27] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Adv. Neural Inf. Process. Syst, 2017, pp. 4765–4774. https://github.com/slundberg/shap. (Accessed 1 January 2021).

[28] T. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (2006) 861–874, https://doi.org/10.1016/j.patrec.2005.10.010.

[29] Q. Liu, D. Li, A.E. Berger, R.A. Johns, L. Gao, Survival and prognostic factors in hypertrophic cardiomyopathy: a meta-analysis, Sci. Rep. 7 (2017) 1–10, https://doi.org/10.1038/s41598-017-12289-4.