Journal of the Serbian Society for Computational Mechanics / Vol. 17 / No. 2, 2023 / pp 153-165
(10.24874/jsscm.2023.17.02.11)

# RISK CLASSIFICATION FOR SUDDEN CARDIAC DEATH IN PATIENTS WITH HYPERTROPHIC CARDIOMYOPATHY BASED ON MACHINE LEARNING ALGORITHMS

**Ognjen Pavić[1,2*], Lazar Dašić[1,2], Tijana Geroski[2,3], Mina Vasković Jovanović[2] and Nenad Filipović[2,3]**

[1]Institute for Information Technologies Kragujevac, University of Kragujevac, Serbia
E-mail: opavic@kg.ac.rs

[2]Faculty of Engineering, University of Kragujevac, Serbia
[3]Bioengineering Research and Development Center (BioIRC), Kragujevac, Serbia
[*]*Corresponding author*

## Abstract

Hypertrophic cardiomyopathy is one of the most prominent cardiovascular diseases, with almost 1 in 500 people suffering from it. It is of great importance for this disease to be detected in a timely manner, so that patients can be provided with an adequate therapy. This is also important for monitoring the future development of the disease so that those patients under a high risk of sudden cardiac death can be provided with lifesaving implantable cardioverter-defibrillators. Regression models were created for the purpose of this paper using the random forest regression algorithm to monitor the future states of patients based on their previously known parameters. Regression models were built by maximizing R2 score for important patient parameters. The training of classification models was done using the random forest and extreme gradient boosted trees algorithms for the purposes of risk prediction. The classification models achieved 96% and 99% F1 score over the high-risk class respectively and 99% prediction accuracy overall.

**Keywords:** hypertrophic cardiomyopathy, extreme gradient boosted trees, random forest, risk prediction, risk classification.

## 1. Introduction

Cardiomyopathy is a generalized name for heart diseases where the walls of the heart muscle are deformed i.e. thickened, stretched or stiffened. Hypertrophic cardiomyopathy (HCM) is a genetic disorder, characterized by the hypertrophy of the left ventricle, which cannot be attributed to secondary sources (Marian and Braunwald, 2017). Left ventricle hypertrophy, that is, the thickened walls of the left ventricle significantly reduce the volume of the atrium and thus the amount of blood the atrium can receive (Keren, Syrris and McKenna, 2008). Thickened walls cannot relax completely, so they become stiffer and more rigid over time. These heart muscle deformations impact the main flow of blood and cause obstructions (Marian and Braunwald, 2017).

In the majority of cases, HCM has a stable course over the years without any major signs of heart failure (HF) (Marian and Braunwald, 2017). However, in some cases, most prominently in

cases in which young adults and adolescents are affected, HCM can be a cause of sudden cardiac death (SCD) Major risk factors of SCD include manifestations of non-sustained ventricular tachycardia, syncope and severe cases of cardiac hypertrophy (Katritsis, Zareba and Camm, 2012). Family history of SCD is also an important indicator that should be consulted for SCD prevention purposes.

High-risk patients cannot be provided optimal protection through only the application of pharmacological therapy, and instead need to receive an implantable cardioverter-defibrillator (ICD) (Gersh et al. 2011). An implantable cardioverter-defibrillator is a device that requires a surgical operation to be implanted subdermally, whose purpose is to monitor the heart rhythm. The device has two electrodes that are introduced into the heart through the right atrium. Through these electrodes, the ICD can deliver a strong electric shock if it detects a faster than normal heartbeat or a series of small electrical shocks if it detects a slower than normal heartbeat, restoring the heart rhythm to a normal pace (heart.org, 2016). Since ICD requires a surgical procedure in order to be implanted and it is permanent, it is of great importance that the ICD is implanted only in patients that are under a lot of danger of dying from complications caused by arrhythmias.

Nevertheless, pharmacological therapy plays its own important role in improving patient quality of life and reducing the risk of further health complications (Ammirati et al. 2016). Pharmacological therapy in HCM is primarily tasked with the control of symptoms, dynamic intraventricular gradient reduction, keeping atrial fibrillation and ventricular arrhythmias in check and preventing cardioembolism (Ammirati et al. 2016).

When assessing cardiomyopathy risk, patient's genetic and clinical features need to be evaluated dynamically, and risk stratification needs to be conducted over a longer period of time. With regard to SCD, based on international guidelines, the high-risk status of a patient has been defined in multiple ways over the years (Christiaans et al. 2010).

Several studies were conducted with the goal of finding the best risk stratification method for HCM patients. Smole et al. (2021) used very similar methodology to ours with a few minor differences. These methodologies differ in the way of utilizing available data for both processes of risk stratification and disease prediction. They achieved classification accuracy of 0.75 and f1 score of 0.71. Kochav et al. (2021) created random forest and gradient boosted decision tree models for risk stratification purposes; however, they concentrated on patients' event data rather than genetic and clinical data. Their evaluation showed the values of 0.88 for sensitivity and 0.84 for specificity metrics. A study was conducted by Augusto et al. (2021) in which left ventricular maximum wall thickness (MWT) was used as the primary feature for risk stratification. Aurore et al. (2018) used mathematical models combined with clustering methods to divide patients into four distinct risk classes. They used the data gathered from HCM patients as well as the data gathered from healthy volunteers for comparison. These data were comprised of genetic data, clinical findings of ECG along with CMR images and extracted T and QRS biomarkers. Tse et al. (2020) used a multilayer perceptron approach to predict the risk of incidents of atrial fibrillation and stroke. Although this study was aimed at predicting heart failure in general, atrial fibrillation prediction can be used for risk stratification of patients with HCM in cases of tachycardia induced cardiomyopathy. Their multilayer perceptron approach yielded results with an f1 score value of 0.88 when assessing incidental atrial fibrillation and 0.87 when assessing transient ischemic attack and stroke, with an f1 score value of 0.89 for all-cause mortality. Wasan et al. (2013) worked on a study which compared machine learning approaches with classical statistical approaches to solve risk stratification problems in predicting heritable arrhythmias.

Although there were several aforementioned studies that deal with implementing patient risk stratification models through the use of machine learning algorithms, most of these models predict the current state of the disease exclusively. Knowledge of the way the disease would be

progressing over time could give doctors a better idea of which therapy should be administered to the patient and at which time. Therefore, we propose a methodology based on machine learning not only to classify patients based on the current state of the disease, but also to predict the future progression of the disease.

## 2. Materials and methods

This part of the paper contains information on the dataset and all steps taken in the process of data preprocessing and the methodology used for data labeling, classification of patients into high-risk and low-risk classes and regression methods used for future prediction. A graphical representation of the methodology we used is shown in Fig. 1.

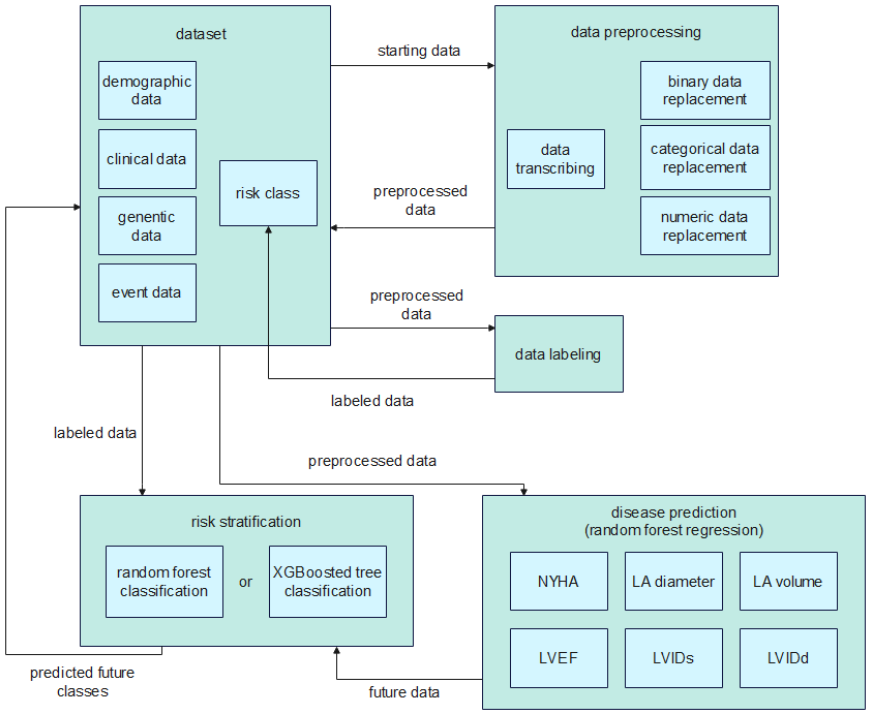


**Fig. 1.** Schematic representation of ML methodology for risk stratification and disease prediction.

### *2.1. Dataset overview*

Our dataset is comprised of 3 distinct subsets of data. The first subset contains patients' personal, clinical and genetic data. These data are gathered from doctor checkups where certain tests are performed. Genetic data are gathered through a blood examination during which the DNA is isolated and certain genes are checked for mutations. During the genetic test, only the genes that are most commonly associated with the development of different types of cardiomyopathies are examined (Columbia cardiology, n.d.). The most common genes associated with HCM are the beta-myosin heavy chain (MYH7) and myosin binding protein C (MYBPC3), which are present in more than 50% of HCM patients (Cahill and Watkins, 2013). Other genes usually tied to HCM are cardiac troponin T (TNNT2), cardiac troponin I (TNNI3), essential myosin light chain

(MYL3), cardiac actin (ACTC1), alpha-tropomyosin (TPM1) and regulatory myosin light chain (MYL2) (Cahill and Watkins, 2013).

For the collection of clinical data, echocardiography and the tissue Doppler tests were used. Echocardiography was used for the purposes of collecting data regarding the heart rhythm and heart muscle shapes. Data required for heart rhythm assessment is gathered from ECG waves and contains information on the length of the P wave, length of the QRS complex and distances between the end of the P wave and the start of the QRS complex and the start of the QRS complex and the T wave end. Data regarding heart muscle shape that can be gathered through echocardiography include intraventricular septum thickness (IVS), posterior wall thickness (PW), left ventricular internal diameter end systole (LVIDs), and at end diastole (LVIDd), left ventricle volume end systole (LVESV) and end diastole (LVEDV). Tissue Doppler test or ultrasound was used for uncovering obstructions of heart valvulae and anomalies in blood circulation. The most important information gained during the tissue Doppler test was the measure of left ventricular outflow tract obstruction (LVOTO) in states of provocation (LVOTO_provocation) and rest (LVOTO_rest). All of the available data points and their descriptions are shown in Table 1.

| Name | Description and possible range of values |
|---|---|
| PersonID | A unique ID number assigned to the patient<br>The number can be any integer value |
| Age | Patient's age<br>An integer value between 10 and 90 |
| Gender | Patient's gender<br>Categorical value [male/female] |
| Primary_Diagnosis | Primary diagnosis of the patient<br>Categorical value [HCM/FAMILY_HISTORY_HCM/DCM/<br>FABRYS_DISEASE/ARVC/HYPERTENSIVE_CARDIOMYOPATHY/<br>CASO_GRIGIO/NO_CLINICAL_FINDING/OTHER] |
| FHx_DCM | Family history of dilated cardiomyopathy<br>Binary value [1 if history of DCM exists, 0 otherwise] |
| FHx_HCM | Family history of hypertrophic cardiomyopathy<br>Binary value [1 if history of HCM exists, 0 otherwise] |
| FHx_SCD | Family history of sudden cardiac death<br>Binary value [1 if history of SCD exists, 0 otherwise] |
| FHx_CAD | Family history of coronary artery disease<br>Binary value [1 if history of CAD exists, 0 otherwise] |
| Alcohol | Note of patient's alcohol consumption<br>Binary value [1 if consumption is present, 0 otherwise] |
| Drug | Note of patient's drug consumption<br>Binary value [1 if consumption is present, 0 otherwise] |
| Pregnancy | Note of the patient's history of pregnancy<br>Binary value[1 if history of pregnancy exists, 0 otherwise] |
| Smoking | Note of the patient's smoking habits<br>Binary value [1 if the habit exists, 0 otherwise] |
| Gene_Testing_Performed | Data point that indicates if genetic test were performed<br>Binary value [1 if test were performed, 0 otherwise] |
| Gene_Name_ACTC1 | Indicates weather ACTC1 has mutations<br>Binary value [1 if a mutation is present, 0 otherwise] |
| Gene_Name_CSRP3 | Indicates weather CSRP3 has mutations<br>Binary value [1 if a mutation is present, 0 otherwise] |
| Gene_Name_MYBPC3 | Indicates weather MYBPC3 has mutations<br>Binary value [1 if a mutation is present, 0 otherwise] |
| Gene_Name_MYH7 | Indicates weather MYH7 has mutations<br>Binary value [1 if a mutation is present, 0 otherwise] |
| Gene_Name_MYL2 | Indicates weather MYL2 has mutations<br>Binary value [1 if a mutation is present, 0 otherwise] |
| Gene_Name_MYL3 | Indicates weather MYL3 has mutations<br>Binary value [1 if a mutation is present, 0 otherwise] |
| Gene_Name_OTHER | Indicates weather a mutation is discovered on a gene that is not usually regarded as a gene of interest<br>Binary value [1 if a mutation is present, 0 otherwise] |
| Gene_Name_PRKAG2 | Indicates weather PRKAG2 has mutations<br>Binary value [1 if a mutation is present, 0 otherwise] |

| | |
|---|---|
| Gene_Name_TNNI3 | Indicates weather TNNI3 has mutations<br>Binary value [1 if a mutation is present, 0 otherwise] |
| Gene_Name_TNNT2 | Indicates weather TNNT2 has mutations<br>Binary value [1 if a mutation is present, 0 otherwise] |
| Gene_Name_TPM1 | Indicates weather TPM1 has mutations<br>Binary value [1 if a mutation is present, 0 otherwise] |
| Gene_Name_TTN | Indicates weather TTN has mutations<br>Binary value [1 if a mutation is present, 0 otherwise] |
| Diagnosis__DIABETES | Patient's diagnosis of diabetes<br>Binary value [1 if present, 0 otherwise] |
| Diagnosis__DIABETES_TYPE_2 | Patient's diagnosis of type 2 diabetes<br>Binary value [1 if present, 0 otherwise] |
| Diagnosis__HYPERCHOLESTEROLEMIA | Patient's diagnosis of hypercholesterolemia<br>Binary value [1 if present, 0 otherwise] |
| Diagnosis__HYPERTENSION | Patient's diagnosis of hypertension<br>Binary value [1 if present, 0 otherwise] |
| Encounter_Date | Data on which the checkup took place<br>Date format: year-month-day |
| Weight | Patient's weight in kilograms<br>Numeric value between 40 and 180 |
| Height | Patient's height in centimeters<br>Numeric value between 100 and 200 |
| BP_Systolic | Value of blood pressure in systole<br>Numeric value between 70 and 190 |
| BP_Diastolic | Value of blood pressure in diastole<br>Numeric value between 30 and 130 |
| NYHA | New York heart association class<br>Numeric value usually between 1 and 4, but can be greater (values above 4 are viewed as if they were 4) |
| BMI | Patient's body mass index<br>Numeric value between 13 and 50 |
| BSA | Bovine serum albumin<br>Numeric value between 0.8 and 2.5 |
| ECG_Inverted_T_Waves | Note of inverted T waves<br>Binary value [1 if T waves are inverted, 0 otherwise] |
| ECG_Pathological_Q_Waves | Note of pathological Q waves<br>Binary value [1 If Q waves are pathological, 0 otherwise] |
| ECG_PR | Distance between the end of P wave and the start of QRS complex<br>Numeric value between 20 and 400 |
| ECG_QRS | Length of the QRS complex<br>Numeric value between 20 and 400 |
| ECG_QTc | Distance between the start of QRS and end of T wave<br>Numeric value between 50 and 1000 |
| ECG_Rate | Distance between two consecutive R waves<br>Numeric value between 30 and 100 |
| ECG_Rhythm | Classification of the ECG response<br>Categorical value [Sinus rhythm/Atrial Fibrillation/Paced/Other] |
| ECG_P | Length of the P wave<br>Numeric value between 10 and 400 |
| Ech_Echo_IVS | Intraventricular septum thickness<br>Numeric value between 5 and 40 |
| Ech_Echo_PW | Posterior wall thickness<br>Numeric value between 5 and 30 |
| Ech_Echo_LA | Left atrium diameter<br>Numeric value between 15 and 80 |
| Ech_Echo_LA_Vol | Left atrium volume<br>Numeric value between 20 and 400 |
| Ech_Echo_Aortic_Root | Aortic root diameter<br>Numeric value between 20 and 50 |
| Ech_Echo_LVEDV | Left ventricular volume end diastole<br>Numeric value between 20 and 300 |
| Ech_Echo_LVEF | Left ventricular ejection fraction<br>Numeric value between 0 and 100 |
| Ech_Echo_LVESV | Left ventricular volume end systole<br>Numeric value between 10 and 230 |
| Ech_Echo_Max_LVT | Left ventricular thrombus<br>Numeric value between 5 and 40 |

| | |
|---|---|
| Ech_Echo_Max_LVT_Loc | Left ventricular thrombus location<br>Categorical value [SEPTUM/POSTERIOR/MID_SEPTUM/<br>APICAL/FREE_WALL/CONCENTRIC/ANTERIOR/OTHER] |
| Ech_Echo_LVIDd | Left ventricular internal diameter (diastolic)<br>Numeric value between 20 and 90 |
| Ech_Echo_LVIDs | Left ventricular internal diameter (systolic)<br>Numeric value between 15 and 70 |
| Ech_doppler_LVOTO_Provocation | Left ventricular outflow tract obstruction in the state of provocation<br>Numeric value between 4 and 180 |
| Ech_doppler_LVOTO_Rest | Left ventricular outflow tract obstruction in the state of rest<br>Numeric value between 3 and 150 |
| Ech_doppler_Mitral_Valve_AVel | Mitral valve velocity A<br>Numeric value between 15 and 180 |
| Ech_doppler_Mitral_Valve_E_DT | E wave deceleration time<br>Numeric value between 30 and 500 |
| Ech_doppler_Mitral_Valve_EVel | Mitral valve velocity E<br>Numeric value between 20 and 200 |
| Ech_tissuedoppler_ave_ea | Average E/A ration<br>Numeric value between 0.1 and 10 |
| Ech_doppler_obstruction | Existence of obstruction<br>Binary value [1 if an obstruction exists, 0 otherwise] |

**Table 1.** Complete dataset overview

The main challenge within this data subset is that not all of the tests are performed at every examination, and that not all patients attended checkups when certain tests were performed. This led to missing data, but had to be replaced in other ways. This subset of data contains binary, categorical and numeric data, and these values were replaced in different ways. The missing data were transcribed from past or future values for the specific patient, where possible. In places where transcribing was not possible, missing data were filled out in one of three ways depending on the type of missing data. Categorical data were replaced by the most numerous categories. Binary data were replaced by choosing values of 1 or 0 so that the distribution stays the same as it was before data preprocessing, while also making sure that the new values are logically possible. Numerical missing data were replaced by the mean of the available values. Additionally, some rows of this subset needed to be dropped due to containing values of certain parameters that were not physically possible, and were presumed to be mistakes when the dataset was being created.

The second data subset contains patients' events and dates when those events occurred. This dataset contained information on patients' history of syncope, which would later be used for risk prediction. The dataset also contained information on events marked as heart failure and sudden cardiac death, which were used to limit which patients' data would be used for prediction of future parameter values for risk evaluation at 5 years. The third data subset contains data on medications prescribed to patients, but this dataset was not used in the final product, because all the relevant data could be gathered from the other two subsets.

### *2.2. Data labeling*

The dataset, after preprocessing, contains 13386 data samples belonging to 3453 distinct patients of which 2178 are male and 1275 are female. These data can be used differently in the processes of risk stratification and the prediction of future values. For risk stratification, every row of the dataset can be interpreted as a possible state of a unique patient, because past values are not required. On the other hand, prediction of future values does require information on the past patient states, so this approach cannot be utilized. The available dataset does not contain class labels, therefore, the first step, after data preprocessing, is labeling every row of available data with a class value of 1 if the patient has a high risk of suffering SCD or 0 otherwise.

It was shown that unsupervised learning was not helpful in grouping the instances. Clustering of the available data was tested using K-means, hierarchical clustering and Gaussian mixture

algorithms, since these algorithms allow the definition of the exact number of expected clusters. However, there is a big overlap between clusters, henceforth, it was crucial to find a different method of labeling data.

Better results with data labeling were achieved when using instructions obtained from medical professionals. These instructions denote 9 conditions of which if 4 are true, the patient would be considered to belong to the high-risk class when attending a checkup.
These conditions include:

- the past diagnosis of syncope,
- New York heart association (NYHA) class>3,
- family history of SCD for patients under the age of 40,
- interventricular septal (IVS) thickness or posterior wall (PW) thickness<30mm,
- left atrium diameter>40mm,
- ejection fraction lower than 50%,
- left ventricular outflow tract pressure gradient (LVOT PG) in resting state>30mmHg,
- N-terminal-pro hormone BNP (NT-proBNP) value greater than 900pg/ml and
- the existence of atrial fibrillation (AF) in any form (Jordà and García-Álvarez, 2018).

After using the proposed method, the results of data labeling show that there exists a large class imbalance. This imbalance needed to be accounted for when training classification models.

### *2.3. Patient classification*

Once the data have been labeled, ML classification models can be trained. Since data clustering could not be applied on unlabeled data the use of kernel based methods like support vector classification is not recommended because of the underlying risk of underfitting the model or bad generalization properties. Deep learning is also not an option in this case, because of the small amount of available data. The main proposition is the use of tree-based ensemble classification models (Boulesteix et al. 2012).

The first model was created using the random forest ensemble algorithm with default parameters. The second model is an extreme gradient boosted tree model created using the XGBoost library. We tested multiple approaches to training these classification models with different features. Namely, we tried training them with only genetic and only clinical data (Smole et al. 2021), however, training the models with both genetic and clinical data together yielded better results. After acquiring these results, tests were ran using scikit-learn and XGBoost inbuilt methods for determining the importance of all features used. Both models showed similar results for certain features, therefore, the final models were trained again using only those features which were graded highly by both methods of measuring importance. Both models have an inherent bias towards predicting the low-risk class, however, the models have great accuracy predicting the high-risk class as well, as shown in the results and discussions section.

## 2.4. Disease prediction

The proposed methodology for future prediction used regression models to predict the future values of certain parameters after which previously trained classification models are used to

predict future patient risk classes. Instructions received from doctors specify 6 parameters that should be monitored over time, in order to monitor progression of the disease.
These parameters are:

- NYHA class,
- left atrium diameter,
- left atrium volume,
- left ventricular ejection fraction,
- left ventricular internal diameter end systole and
- left ventricular internal diameter end diastole.

Random forest based regression models were trained by utilizing other patient clinical data, along with the previous values of proposed parameters. The criteria by which the training data for these regression models were chosen are based on the number of appearances of unique patients within the dataset. Namely, the patient needed to have at least two doctor visits during which the observed parameter was measured, not including imputed parameter values. Additionally, patients who were marked as having suffered a sudden cardiac death or heart failure in the event subset were not used for training due to the perceived impossibility of obtained future parameter values. Afterwards, the new data were compiled by transcribing patients' old clinical data and feeding them into proper regression models. The amount of future data samples added into the future risk classification dataset was based on the average amount of time passed between checkups for those patients. Patients who had only a single recorded doctor visit always had 5 new data samples added into the future disease classification dataset in one year increments. Patients who had suffered HF or SCD did not receive additional data samples in the future disease classification dataset because the predicted data could not be used in a productive manner. New data were classified, so that disease progression and the associated risk class could be monitored over time.

## 3. Results and discussion

In the scope of this study, we have developed multiple machine learning models for the purposes of risk classification and prediction of disease progression over the next 5 years. This part of the paper is dedicated to evaluating the acquired results as well as comparisons to previously conducted studies in the same field.

### *3.1. Classification results*

The first goal of the paper was the classification of patients suffering from hypertrophic cardiomyopathy into high-risk and low-risk classes in terms of risk of SCD. Both of the proposed classification models have very good classification accuracy as shown in Fig. 2a for random forest and Fig. 2b for extreme gradient boosted decision tree.

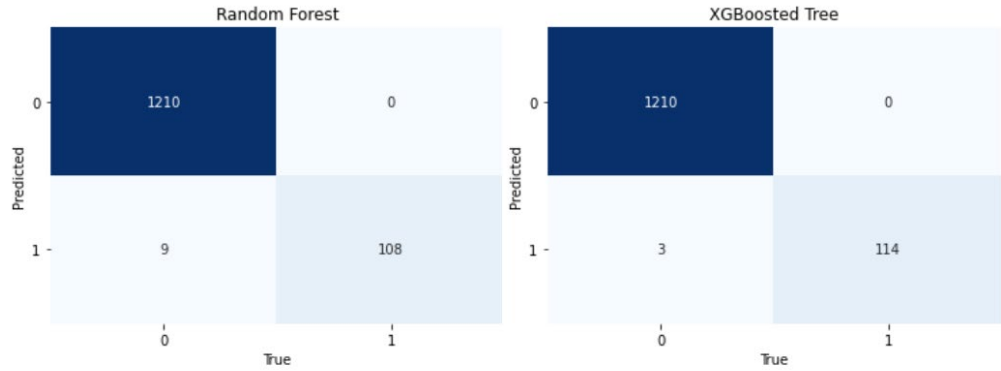


**Fig. 2.** Confusion matrix for (a) random forest classification model and (b) XGBoosted decision tree model.

However, prediction accuracy is not a relevant evaluation metric for these models because of data imbalance. High-risk class is much smaller than the low-risk class so there may exist an inherent prediction bias towards the larger class. Also, it is imperative to predict the high-risk class correctly in as many cases as possible due to the consequences of incorrect prediction in these cases. Therefore, F1 score for the high-risk class was used as the main evaluation metric for our models. A complete overview of achieved classification sensitivity, specificity, precision, negative predictive value (NPV), F1 score and prediction accuracy metrics are presented in Table 2.

| | Random Forest | | XGBoosted trees | |
|---|---|---|---|---|
| | Low Risk (0) | High Risk (1) | Low Risk (0) | High Risk (1) |
| Sensitivity | 0.99 | 1 | 1 | 1 |
| Specificity | 1 | 0.99 | 1 | 1 |
| Precision | 1 | 0.92 | 1 | 0.97 |
| NPV | 0.92 | 1 | 0.97 | 1 |
| F1-score | 1 | 0.96 | 1 | 0.99 |
| Accuracy | 0.99 | 0.99 | 0.99 | 0.99 |

**Table 2.** Evaluation metrics of classification models

Both models achieved very similar results, with the XGBoosted tree model only being slightly better than the model created using the random forest algorithm, making them interchangeable in the future prediction step of our research.

### *3.2. Regression metrics*

Regression models were created with a goal of predicting future states of important patient parameters that can be used in evaluating disease progression over time. Each of these parameters required the training of a separate regression model. Separate models were built with regards to maximizing the R2 score metric to insure the highest possible correlation between the most important patient parameters that the system is predicting and other data points. These models were evaluated using the mean squared error metric. Results are denoted in Table 3.

| | R2 score | Mean squared error |
|---|---|---|
| NYHA | 0.56 | 0.22 |
| LA | 0.76 | 16.88 mm |
| LAvol | 0.8 | 35.11 $mm^3$ |
| LVIDd | 0.69 | 13.19 mm |
| LVIDs | 0.81 | 7.8 mm |
| LVEF | 0.7 | 33.23 % |

**Table 3.** Evaluation metrics of different regression models

Mean squared error metrics are compared with the span of possible values for each parameter to evaluate the accuracy of our regression models. Regression models show very good prediction properties except for the NYHA class regression model. In future research, regression models will be created using additional data on applied pharmacological therapies and their impact on the patient over time. However, this inclusion will require data on disease progression over a longer period of time for a larger set of patients.

### *3.3. Discussion*

Although there were multiple studies conducted on this subject, most of them do not have the same approach to solving the problem at hand. Most of the studies that use images as risk stratification data suffer from the inability to predict future disease progression. Also, in many of the cases, even though other types of data were used, the study was focused strictly on classifying the current state of the disease. In this study, we strived towards predicting the disease progression over multiple years as well as the current state so that pharmacological therapy could be considered in advance, before the symptoms became too severe for intervention.

Smole et al. (2021) had a different approach to using the given dataset, wherein they did not view each patients visit to an examination as a separate entity, but viewed all of the single patients' visits as a single entity. Therefore, using the proposed methodology we had more samples for training and testing our classification models and achieved better results in terms of f1 score. Both of our classification models outperformed those presented in the study of Kochav et al. (2021) who used event centric data for risk stratification. Aurore, et al. (2018) achieved great results with their approach to solving this problem, however, we are unable to make a logical comparison between the two proposed methodologies, due to the uncertainty of the utilization of unsupervised learning methods in the process of classification and also the discrepancies between the classes that were being predicted. Tse et al. (2020) created models using machine learning approaches, for risk stratification of HF in patients suffering not only from HCM, but other heart complications as well. However, our model achieved better accuracy for specifically SCD caused by HCM. The comparison between similar studies is shown in Table 4. Attention needs to be drawn to the fact that the performance indicator varies between these studies because of the differing approaches and decisions made by authors.

| | (Smole et al. 2021) | (Kochav et al. 2021) | (Aurore et al. 2018) | (Tse et al. 2020) | Proposed methodology |
|---|---|---|---|---|---|
| Investigated problem | HCM patient stratification on clinical data | HCM patient stratification on event data | ECG phenotypes in HCM patients | Stratification of heart failure | HCM stratification on clinical and genetic data |
| Binary/Multi classification | Binary (low /high risk) | Binary (low/high risk) | Multi (healthy, low/medium/high risk) | Binary (low/high risk) | Binary (low/ high risk) |
| Performance indicator and value | High risk F1 score: 0.71 | Sensitivity: 0.88 | Not applicable | High risk F1 score: 0.89 | High risk F1 score: 0.99 |

**Table 4.** Overview of results in similar studies

While our final evaluation shows great promise in predicting the disease progression of hypertrophic cardiomyopathy, it can still be speculated about the improvements of these results in the future. One of the main limitations of this research was the amount of data available in the dataset, although it was sufficient for classification purposes, many patients had only one or two recorded sets of parameters weighing down the possibility of obtaining better future parameter value prediction results through regression. One of the proposed methods of enhancing this dataset is creating more patient data through finite element method physical simulations. In this way, we could obtain more data, much faster than it could be obtained otherwise, using physical laws that govern the heart muscle contractions and blood flow, therefore improving prediction methods to be even more precise than they currently are.

## 4. Conclusion

The main goal of this research was not to replace medical professionals in diagnosing HCM in patients, but to provide a decision support system whose purpose is to lend a helping hand in examining large amounts of data and providing a second opinion. The current gold standard for diagnosing hypertrophic cardiomyopathy is a calculator that is able to predict the future state of HCM patients in the next 5 years (O'Mahony et al. 2014). While this calculator already exists, it is based on a concrete mathematical equation that reports the patient's risk class in different time frames. This study was conducted using new data that were not available in the creation of the HCM calculator. Moreover, during the creation of this system, it was imperative to predict future values of many different data points and not only the risk class. These crucial future data points are also available to medical professionals for further inference into solving the task at hand. Additionally, the decision making nature of the utilized tree based classification models can outperform the existing mathematical equation in edge cases when data is extrapolated, due to good generalization properties.

In conclusion, it was possible to create a decision support system that predicts the current state of the patient almost perfectly while also having a high amount of success in predicting the future development of the disease. The regression section of the system was built by maximizing R2 score for important patient parameters that ultimately amounted to 81% LVIDs, 69% LVIDd, 80% LAvol, 70% LVEF, 76% LA and 56% NYHA. The classification section of the system has 96% or 99% f1 score over the high-risk class depending on if the algorithm in question is random forest or XGBoost and 99% prediction accuracy over all. The system can be improved further through gathering additional data. Likewise, in future endeavors, the scope of this research will be expanded to focus on the creation of a decision explanation module whose purpose would be

to give doctors an insight into the inner workings of the system and raise the trust of patients towards an automated diagnostics process.

**Acknowledgements:** This paper has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101080905. (STRATIFYHF project). The research is supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, contract numbers [451-03-47/2023-01/200107 (Faculty of Engineering, University of Kragujevac) and 451-03-47/2023-01/200378 (Institute for Information Technologies, University of Kragujevac)]. This research is also supported by the project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777204. (SILICOFCM project). This article reflects only the author's view. The Commission is not responsible for any use that may be made of the information it contains.

## References

Ammirati, E., Contri, R., Coppini, R., Cecchi, F., Frigerio, M., and Olivotto, I. (2016, September). Pharmacological treatment of hypertrophic cardiomyopathy: current practice and novel perspectives. *European journal of Heart Failure, 18*(9), 1106-1118.

Augusto, J. B., Davies, R. H., Bhuva, A. N., Knott, K. D., Seraphim, A., Alfarih, M., Ntusi, N. A. (2021). Diagnosis and risk stratification in hypertrophic cardiomyopathy using machine learning wall thickness measurement: a comparison with human test-retest performance. *The Lancet Digital Health, 3*(1), e20-e28.

Aurore, L., Rina, A., Ana, M., Masliza, M., Elizabeth, O., Pablo, L., Blanca, R. (2018). Distinct ECG Phenotypes Identified in Hypertrophic Cardiomyopathy Using Machine Learning Associate With Arrhythmic Risk Markers. *Frontiers in Physiology, 9*.

Boulesteix, A.-L., Janitza, S., Kruppa, J., and König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*(6), 483-507.

Cahill, T. J., and Watkins, H. A. (2013). Genetic Cardiomyopathies Causing Heart Failure. *Circulation Research, 113*(6).

Christiaans, I., Engelen, K. v., Langen, I. M., Birnie, E., Bonsel, G. J., Elliott, P. M., and Wilde, A. A. (2010, March). Risk stratification for sudden cardiac death in hypertrophic cardiomyopathy: systematic review of clinical risk markers. *EP Europace, 12*(3), 313-321.

*Columbia cardiology*. (n.d.). (Columbia university, department of cardiology) Retrieved 8. 12., 2022., from https://www.columbiacardiology.org/patient-care/hypertrophic-cardiomyopathy-center/about-hypertrophic-cardiomyopathy/genetic-testing

Gersh, B. J., Maron, B. J., Bonow, R. O., Dearani, J. A., Fifer, M. A., Link, M. S., Yancy, C. W. (2011). 2011 ACCF/AHA Guideline for the Diagnosis and Treatment of Hypertrophic Cardiomyopathy: Executive Summary. *Circulation*, 2761–2796.

*heart.org*. (2016, 9 30). (.) Retrieved 7 20, 2022, from https://www.heart.org/en/health-topics/arrhythmia/prevention--treatment-of-arrhythmia/implantable-cardioverter-defibrillator-icd

Jordà, P., and García-Álvarez, A. (2018, August 12). Hypertrophic cardiomyopathy: Sudden cardiac death risk stratification in adults. *Global cardiology science and practice*.

Katritsis, D. G., Zareba, W., and Camm, A. J. (2012 , November). Nonsustained Ventricular Tachycardia. *Journal of the American College of Cardiology, 60*(20).

Keren, A., Syrris, P., and McKenna, W. J. (2008, January). Hypertrophic cardiomyopathy: the genetic determinants of clinical disease expression. *Nature Reviews Cardiology, 5*.

Kochav, S. M., Raita, Y., Fifer, M. A., Takayama, H., Ginns, J., Maurer, M. S., Shimada, Y. J. (2021). Predicting the development of adverse cardiac events in patients with hypertrophic cardiomyopathy using machine learning. *International Journal of Cardiology, 327*, 117-124.

Marian, A. J., and Braunwald, E. (2017). Hypertrophic Cardiomyopathy: Genetics, Pathogenesis, Clinical Manifestations, Diagnosis, and Therapy. *Circulation research, 121*, 749–770.

O'Mahony, C., Jichi, F., Pavlou, M., Monserrat, L., Anastasakis, A., Rapezzi, C., McKenna, W. J. (2014). A novel clinical risk prediction model for sudden cardiac death in hypertrophic cardiomyopathy (HCM Risk-SCD). *European Heart Journal, 35*(30), 2010-2020.

Smole, T., Žunkovič, B., Pičulin, M., Kokalj, E., Robnik-Šikonja, M., Kukar, M., MacGowan, G. A. (2021). A machine learning-based risk stratification model for ventricular tachycardia and heart failure in hypertrophic cardiomyopathy. *Computers in Biology and Medicine, 135*.

Tse, G., Zhou, J., Woo, S. W., Ko, C. H., Lai, R. W., Liu, T., Zhang, Q. (2020). Multi-modality machine learning approach for risk stratification in heart failure with left ventricular ejection fraction ≤ 45%. *ESC Heart Failure, 7*(6), 3716-3725.

Wasan, P., Uttamchandani, M., Moochhala, S., Yap, V., and Yap, P. (2013). Application of statistics and machine learning for risk stratification of heritable cardiac arrhythmias. *Expert Systems with Applications, 40*(7), 2476-2486.