The Importance of Genetic and Clinical Data Features in Risk Stratification of Patients with Hypertrophic Cardiomyopathy

Ognjen Pavić[1,2][0000-0003-2533-1079], Lazar Dašić[1,2][0000-0002-8055-100X], Tijana Geroski[2,3][0000-0003-1417-0521], Anđela Blagojević[3][0000-0002-8652-3827] and Nenad Filipović[2,3][0000-0001-9964-5615]

[1] Institute of Information Technologies Kragujevac, University of Kragujevac, Kragujevac, Serbia
[2] Bioengineering Research and Development Center (BioIRC), Kragujevac, Serbia
[3] Faculty of Engineering, University of Kragujevac, Kragujevac, Serbia
opavic@kg.ac.rs

Abstract. Hypertrophic cardiomyopathy is a genetic cardiovascular disease which affects the heart's left ventricle. This paper presents the results obtained from examining the importance of clinical and genetic data points in risk stratification of patients with hypertrophic cardiomyopathy. The significance of features was gathered in consultations with cardiologists as well as from the evaluation of created classification models built for the purposes of risk assessment. The main goal of the study was to find hidden knowledge within the dataset that could be used to further improve classification results and to compare the aforementioned knowledge with the information gathered from doctors with the goal of potentially improving the manual diagnostics approach. The study was conducted on genetic and clinical data separately as well as on a combined dataset. The importance of parameters was calculated with two different classification models, and was also calculated using two different methods of manual data annotation. All of the acquired results show both similarities and differences from one another. The acquired results were evaluated based on the predictive abilities of classification models.

Keywords: cardiomyopathy, machine learning, data importance, diagnostics, biomarker analysis, classification, risk assessment

1 Introduction

Hypertrophic cardiomyopathy (HCM) is a genetic cardiovascular disorder that is characterized by the hypertrophy of the heart muscle walls of the heart's left ventricle [1]. The hypertrophy, or thickening of the left ventricle walls affects the stiffness and the rigidity of the heart muscle tissue as a consequence of the fact that thickened walls cannot relax properly during the cardiac cycle. Hypertrophy of the left ventricle walls directly impacts the blood flow through the heart and causes obstructions. In the majority of cases, HCM has a stable course over the years without any major complica-

tions; however it is essential for HCM to be properly diagnosed on time because it can lead to the development of arrhythmias, heart failure, stroke and death [1].

When calculating the severity of HCM, the main classification target is the risk of suffering a sudden cardiac death. Sudden cardiac death is a death attributed to a cardiovascular cause which happens within one hour after the onset of symptoms [2].

HCM can be diagnosed through genetic testing as well as through echocardiography; therefore this paper aims to calculate the importance of features for classifying patients into high-risk and low-risk classes based on the risk of suffering a sudden cardiac death using genetic test findings, clinical echocardiography findings, and both methods in tandem with one another.

Most of the available research [3,4,5] focuses mainly on classifying patients into two risk classes with regards to the risk of suffering a sudden cardiac death and glosses over feature importance, viewing it only as a tool for minor prediction accuracy improvement. In their paper, Smole et al. [3] used a very similar methodology for patient classification to the methodology that was employed during this study, with differences arising in the methods of data sample utilization. Namely, in their research, each patient was viewed as a single entity, while our study views each visit to the doctor as a possible state that any patient can find themselves in, at some point in time and thus utilizes each visit as a separate entity, thereby creating a larger number of training and testing samples for resulting classification models. In contrast to their approach, Kochav et al. [4] used random forest and extreme gradient boosted trees algorithms for patient stratification, however their models were trained using mainly data on events that have happened to patients in the past. A study was conducted by João et al. [5] in which left ventricular maximum wall thickness (MWT) was used as the primary feature for risk stratification. Aurore et al. [6] used mathematical models combined with clustering methods to divide patients into four distinct risk classes. They used data gathered from healthy volunteers as well as HCM patients for comparison. This data was comprised of a genetic dataset and a clinical dataset which contained ECG along with CMR images and extracted T and QRS biomarkers. In their study, Tse et al. [7] utilized a multilayer perceptron approach to solving a risk stratification problem regarding incidents induced by atrial fibrillation and stroke. While their study was aimed at cardiovascular diseases in general, the stratification of risk of death amidst atrial fibrillation includes hypertrophic cardiomyopathy into the disease interest group. Even though each of the aforementioned studies achieved satisfactory classification results in their own rights, they all lack a certain degree of result explanation as well as decision making process explainability. With a deeper assessment of feature importance for different feature sets, our approach to risk stratification provides additional insight into both the achieved results and the way classification models make decisions. The assessment of the significance of certain features contained in the available dataset also provides a baseline for future improvement in terms of providing stable grounds for the construction of explanation modules aimed at explaining results in a more detailed way to patients and medical professionals alike.

On the other hand, there exists medical research [8] specifically aimed at discovering new ways of diagnosing heart diseases like HCM. This paper aims to utilize the AI centered approach to risk stratification while also paying close attention to expla-

nations of machine learning models regarding the way they learn and classify given data. This approach opens the possibilities of discovering hidden knowledge in the data that can improve the manual diagnostics process while also providing an automated machine learning based model that has the possibility to serve as a decision support tool.

2 Methodology

The dataset is comprised of demographic, genetic and clinical data as well as clinical investigations and disease related events. The dataset contains a total of 13386 samples, collected from 2302 distinct patients, gathered during multiple attended checkups. The retrospective data that was used for training the machine learning algorithms was provided by the Careggi University Hospital, University of Florence, Italy. Retrospective data was gathered over a 13 month period, during which clinical test were performed in regular intervals. ECG and Doppler tests were performed during months 1 and 12, Holter test were performed during months 2 and 13, while CMRI findings were gathered during month 6. The inclusion criteria for patients were a primary diagnosis of HCM or the existence of a HCM diagnosed relative.

The available dataset first needed to be processed and brought into a state usable for classification model training. The set contained instances of missing data that was filled in using transcription of past or future values. In cases in which transcription of data was not possible, missing data was filled in using other common data imputation techniques. Namely missing numeric data was filled in using the mean value of the observed variable; missing categorical data was filled in using the category which is most numerous. When it comes to binary data imputation, a system was devised to input values of 0 or 1, while paying attention not to assign different values of binary variables to the same patient when filling in data missing in follow-ups and while also paying close attention to the distribution of new values so that the distribution stays the same after imputation as it was before imputation. The dataset also contained extreme, physically impossible values that were eliminated from the set before model training. Finally, none of the patient data in the available dataset was labeled with risk classes, so labeling had to be conducted as a way of creating the classification target.

The first approach to data labeling was using doctors' instructions. Cardiologists named the following 9 criteria:

1. Past diagnosis of syncope
2. New York heart association (NYHA) class value greater than 3
3. Family history of sudden cardiac death while the patient is younger than 40 years of age
4. Interventricular septum (IVS) thickness or posterior wall (PW) thickness less than 30mm
5. Left atrium diameter greater than 40mm
6. Ejection fraction lower than 50%

7. Left ventricular outflow tract pressure gradient (LVOT PG) in resting state higher than 30mmHg
8. N-terminal-pro hormone BNP (NT-proBNP) value greater than 900pg/ml
9. The existence of atrial fibrillation in any form

Of these 9 criteria if 4 or more were true, the patient is classified as having a high risk of suffering a sudden cardiac death.

The second approach to data labeling was using the information on disease related events. If an event corresponding to high risk of sudden cardiac death occurred in the event dataset, the patient would be labeled as high risk from that point in time onwards. Events that were taken into account as being closely related to having a high risk of suffering a sudden cardiac death were the following:

1. Arrhythmia – non-sustained ventricular tachycardia (NSVT)
2. Arrhythmia – sustained ventricular tachycardia (SVT)
3. Abnormal Holter
4. Abnormal exercise tolerance test (ETT)
5. Heart failure

Additionally the high-risk class was attributed to patients who were marked as dead from suffering a sudden cardiac death, and also patients who had an implantable cardioverter defibrillator, readings gathered from an implantable cardioverter defibrillator, had a heart transplant or were marked to receive a heart transplant.

Using these two approaches, two datasets were created for later comparison, for both feature importance and model prediction accuracy.

For the purposes of classification, two different ensemble classification groups of models were created. The first group was built using the random forest algorithm, while the second group was created using the extreme gradient boosted trees method. All of these models were trained and evaluated using 6 different datasets. Namely, each model was trained using only genetic data, using only clinical data or using both genetic and clinical data together. Training was conducted with both labels created from cardiologists' instructions and those created from disease related event data. In the end, there were 12 results gathered from distinct combinations of inputs, outputs and model creation algorithms. The entire methodology diagram is shown in Figure 1.

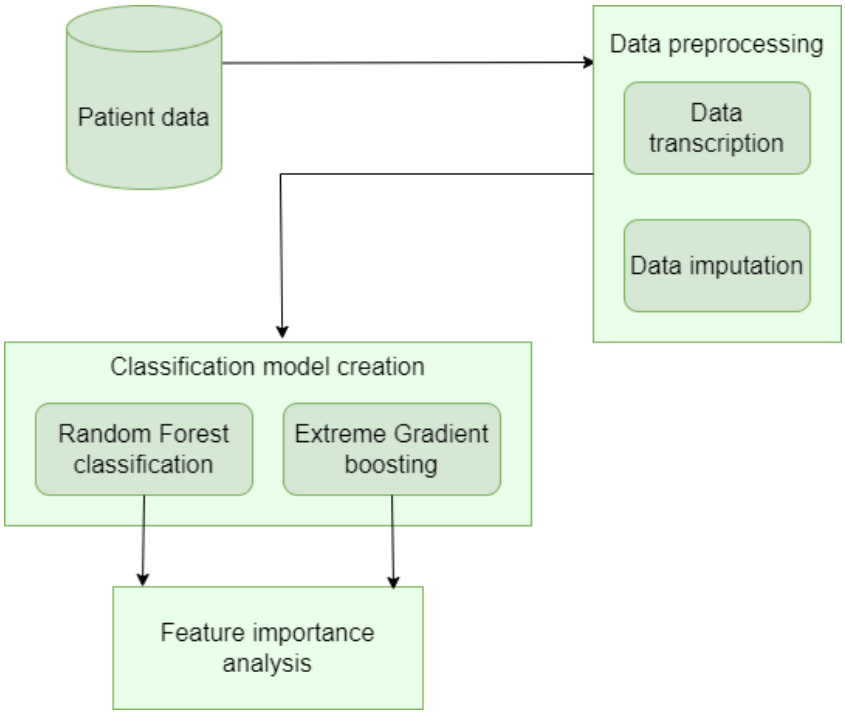

Fig. 1. Patient classification and feature importance analysis methodology diagram

Data importance was calculated for each of the resulting 12 classification models. It is important to note that not all of the models achieved good classification accuracy, therefore when assessing the importance of certain data features, prediction metrics were also taken into account.

3 Results and discussion

After conducting classification using all 12 combinations of inputs, labeling methods and classification algorithms, the models were evaluated based on prediction accuracy. The achieved results are shown in Table 1 for random forest-based models, as well as in Table 2 for extreme gradient boosted trees-based models.

Table 1. Prediction accuracy for random forest classification models

	Gold standard (doctor)	Event labeling
Genetic data	90.56%	60.98%
Clinical data	92.66%	92.55%
Combined data	94.37%	94.21%

Table 2. Prediction accuracy for extreme gradient boosted trees classification models

	Gold standard (doctor)	Event labeling
Genetic data	90.52%	60.92%
Clinical data	92.18%	91.64%
Combined data	96.69%	96.53%

The following conclusions were drawn from these results. Both approaches to creating classification models are very close in terms of prediction accuracy, with extreme gradient boosting exceling in full dataset classification, while the random forest is better at classification using individual parts of the dataset. Training the models using data labeled through disease related events is worse no matter the input data used, especially when training models using genetic data. With these results, data importance for different sets of inputs was valued closely between models and data labeling approaches, except the importance calculated for genetic data with event labeling which was completely ignored due to the extremely poor results.

Table 3 shows the most important clinical data features. There exist some variations in most important clinical features between random forest (RF) and extreme gradient boosted trees (XGB) algorithms, so only features which were of high importance in every training case were chosen. The most prominent clinical features include left atrium (LA) volume, LA diameter, interventricular septum thickness (IVS), left ventricular ejection fraction (LVEF), left ventricular internal diameter end systole (LVIDs), left ventricular internal diameter end diastole (LVIDd), left ventricular end systole volume (LVESV), left ventricular end diastole volume (LVEDV), ECG rhythm shape and n-terminal pro hormone BNP (NTBNP) concentration.

It is important to note that, even though it does not play the most important role in decision making while training random forest classification models, ECG rhythm plays by far the most important role when training extreme gradient boosted trees classification models, with it being responsible for slightly more than 15% of the decision making process.

Table 3. Feature importance of clinical data

	Feature importance			
	Gold standard labeling		Event labeling	
Feature name	RF	XGB	RF	XGB
LA volume	6.2%	3.9%	5.2%	3.9%
LA diameter	5.4%	5.6%	5.6%	5.6%
IVS	3.9%	3.2%	3.8%	3.2%
LVEF	3.7%	3.3%	3.4%	3.3%
LVIDs	2.3%	1.3%	2.4%	1.3%
LVIDd	2.1%	1.3%	2.1%	1.3%
NTBNP	2.3%	2.7%	2.3%	2.7%
ECG Rhythm	2%	15.1%	1.8%	15.1%
LVEDV	2%	1.6%	2.1%	1.6%
LVESV	1.9%	1.6%	2.1%	1.6%

Table 4 shows the most important genetic data features which were evaluated as having a high degree of importance for both classification model creation algorithms. From the following results it can be seen that the importance of genes varies greatly between classification algorithms, especially in the case of MYL3 and TPM1 genes. However in both cases MYBPC3 and ACTC1 genes are responsible for making the greater part of the decision during classification, making up nearly 52% of the decision making process for extreme gradient boosted trees classification and 77% of the decision making process for random forest classification.

Table 4. Feature importance of genetic data

	Feature importance	
Feature name	RF	XGB
MYBPC3	66.9%	29.6%
ACTC1	10.3%	22%
TNNI3	6%	5.8%
TNNT2	1.7%	3%
MYL3	1%	12%
MYL2	1.6%	1.9%
TPM1	1.9%	11.3%

Table 5 shows the importance of features within in the grand scheme when classification is conducted using all of the available data. The importance of most prominent features is more stable across the board when training classification algorithms with the exception of ECG rhythm which still plays a disproportionately more important role for the decision making process of extreme gradient boosted tree classifiers. It is important to note that the most important feature importance values are lower when using the entire available dataset because of the increase in the number of features used for training as well as the more balanced role each feature plays in the decision making process.

Table 5. Feature importance of the full dataset

	Feature importance			
	Gold standard labeling		Event labeling	
Feature name	RF	XGB	RF	XGB
LA volume	3.7%	2.6%	4.3%	2.6%
LA diameter	5.1%	4%	4.4%	4%
IVS	3.2%	2.5%	3%	2.5%
LVEF	2.5%	2.7%	2.6%	2.7%
ECG rhythm	2.3%	7.8%	2.2%	7.8%
MYBPC3	2.8%	4.6%	2.8%	4.6%
NYHA	2.1%	4.6%	2.1%	4.6%
Age	2.6%	2.3%	2.7%	2.3%

The classification models which were trained using the entire available dataset achieved higher prediction accuracy than models which were trained using only clinical or only genetic data. The most notable conclusion that can be drawn from the final results are that clinical data plays a much bigger role in the decision making process than genetic data and that demographic data which was included only in the decision making process of models trained using the entire dataset also plays a big role in achieving accurate classification, most prominently patient age and New York heart association class.

4 Conclusion

Although there were multiple studies conducted on the subject of patient risk stratification for risk of suffering a sudden cardiac death caused by hypertrophic cardiomyopathy, none of those studies focus on uncovering the importance of features used for said stratification. In many of those cases satisfactory results are achieved but are not elaborated upon. When tackling problems of this nature it is important to have a degree of explainability to both further the knowledge on the subject matter and increase the likelihood of the created technology to be adopted by medical professionals and patients alike.

Our classification models for risk stratification achieved great results especially in the case when the entire feature set was used during model training. When it comes to the explainability of the models, feature importance calculation, although not the only approach, provides a deeper insight into the inner working of the developed models thereby making the utilized black box approaches more see through.

In order to further improve the presented models in the future, we plan to train new classification models, which will be trained using best combinations of patient attributes based on the discovered significance of said attributes. Additionally, an explanation module is planned to accompany the improved classification models which will make the decision making process even more understandable to both patients and medical professionals potentially increasing the degree of trust in the system.

Acknowledgements

The research was funded by the project that has received funding from the European Union's Horizon 2020 research and innovation programmes under grant agreement No 952603 (SGABU project). This paper is supported by the SILICOFCM project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777204. This research is also supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, [Contract No. 451-03-47/2023-01/200378, (Institute for Information Technologies Kragujevac, University of Kragujevac)]. 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

1. B.J. Maron, M.S. Maron, (2013), *Hypertrophic cardiomyopathy*, Lancet, Vol. 381, 9862. pp. 242-255
2. R. Virmani, A.P. Burke, A. Farb, (2001), S*udden Cardiac Death, Cardiovascular Pathology*, Vol. 10, 5. pp. 211-218
3. T. Smole et al. (2021), *A machine learning-based risk stratification model for ventricular tachycardia and heart failure in hypertrophic cardiomyopathy*, Computers in Biology and Medicine, vol. 135.
4. S.M. Kochav et al., (2021), *Predicting the development of adverse cardiac events in patients with hypertrophic cardiomyopathy using machine learning*, International Journal of Cardiology, Vol. 327. pp. 117-124
5. B. A. João et al. (2021), *Diagnosis and risk stratification in hypertrophic cardiomyopathy using machine learning wall thickness measurement: a comparison with human test-retest performance*, Lancet Digital Health. 3(1): pp. 20-28.
6. L. Aurore et al. (2018), *Distinct ECG Phenotypes Identified in Hypertrophic Cardiomyopathy Using Machine Learning Associate With Arrhythmic Risk Markers*, Frontiers in Physiology.; 9.
7. Gary Tse et al. (2020), *Multi-modality machine learning approach for risk stratification in heart failure with left ventricular ejection fraction ≤ 45*, National Library of Medicine, pp. 3716-3725
8. E. L. Matthia et al., (2022), *Circulating Biomarkers in Hypertrophic Cardiomyopathy*, Journal of American Heart Association.