Separation of HCM and LQT cardiac diseases with machine learning of Ca^{2+} transient profiles

Henry Joutsijoki[1], Kirsi Penttinen[2], Martti Juhola*[1] and Katriina Aalto-Setälä[2,3]

[1]Faculty of Information Technology and Communication Sciences, [2]Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland, [3]Heart Center, Tampere University Hospital, Tampere, Finland

*Correspondence: Martti.Juhola@tuni.fi

Abstract

Background: Modelling human cardiac diseases with induced pluripotent stem cells (iPSCs) enables to study disease pathophysiology and to develop therapies but also, as we have previously showed, it can offer a tool for disease diagnostics. We observed previously that a few genetic cardiac diseases can be separated from each other and healthy controls by applying machine learning to Ca^{2+} transient signals measured from iPSC-derived cardiomyocytes (CMs).

Data and methods: For the current research, 419 hypertrophic cardiomyopathy (HCM) transient signals and 228 long QT syndrome (LQTS) transient signals were measured. HCM signals included data recorded from iPSC-CMs carrying either TPM1 (HCMT) or MYBPC3 (HCMM) mutation and LQTS signals included data recorded from iPSC-CMs carrying KCNQ1 mutation (LQT1) or KCNH2 mutation (LQT2). After preprocessing those Ca^{2+} signals where we computed peak waveforms we classified the two mutations of both disease pairs by using machine learning methods.

Main finding: We obtained excellent classification accuracies of 89% for HCM and even 100% for LQT at their best.

Principal conclusion: The results indicate that the methods applied would be efficient for the identification of these genetic cardiac diseases.

Keywords: Calcium transient profiles, genetic cardiac diseases, machine learning, separation of mutations

1. Introduction

Induced pluripotent stem cell-derived cardiomyocytes (iPSC-CMs) have enabled the study of different genetic cardiac diseases like long QT syndrome (LQT) [1-4] and hypertrophic cardiomyopathy (HCM) [5-7]. Functional studies of cardiomyocyte calcium cycling have revealed new insights into different cardiac diseases by showing substantial defects and abnormalities in the calcium cycling of iPSC-CMs, reflecting the cardiac phenotype observed in patients [1,8]. Calcium cycling has a central role in cardiac functionality by linking electrical activation and contraction, and therefore it is important to investigate it to improve the studies of disease pathology, prevention and treatment but also, as we have shown earlier [9], in disease diagnostics.

Hypertrophic cardiomyopathy (HCM) is a genetic cardiac disease, which affects the structure of heart muscle tissue and can lead to arrhythmias and a heart failure [6]. The majority of the mutations are found either in α-tropomyosin (TPM1) of the β-myosin heavy chain (MYH7) or in the myosin-binding protein C (MYBPC3) genes [10] and these mutations are called HCMT and HCMM, respectively. Long QT syndrome (LQTS) is a potentially severe arrhythmic disease that affects the electrical repolarization of the myocardium and manifests as an abnormally long QT interval on electrocardiogram recordings. In LQT type 1 (LQT1), the mutations are present in the KCNQ1 gene, which encodes the α-subunit of the slow component of the delayed rectifier potassium current (Kv7.1) [11]. LQT2 is caused by mutations of the KCNH2 gene (also known as human ether-a-go-go-related gene (HERG)) which encodes the α-subunit of the rapid component of the delayed rectifier potassium current (Kv11.1) [12].

Thus far, machine learning has still seldom been utilized for data collected from iPSC-CMs. Mechanistic action of cardioactive drugs has been investigated [13]. We found out in our previous research [14,15] that abnormally and normally beating cardiomyocytes can be differentiated from each other. Furthermore, we observed that three genetic cardiac diseases including jointly HCM and LQT1 and catecholaminergic polymorphic ventricular tachycardia (CPVT) can be separated from healthy controls (wild type, WT) by classifying the Ca^{2+} transient signals of the diseased as one class and those of the controls as the other [16]. We showed that those three diseases can be separated from each other and each of them from the cardiomyocytes of the healthy controls [9], where we also discovered that it was not necessary to consider abnormally and normally beating cardiomyocytes within each group as

separate groups. They could be classified jointly, since both groups could be separated approximately equally well between three diseases and controls. This is extremely advantageous for practical reasons when thinking of future research and even potential clinical applications.

In the current research, two mutations of HCM disease including HCMM and HCMT and two types of LQTS including KCNQ1 gene mutation (LQT1) or the KCNH2 (HERG) gene mutation (LQT2) [6] were studied in order to separate between two types of each disease. Several different machine learning algorithms were run with Matlab and experimented with the present Ca^{2+} transient signal data of the iPSC cardiomyocytes.

2. Material

The study was approved by the Ethics Committee of Pirkanmaa Hospital District in establishing, culturing and differentiating human iPSC lines (R08070). Patient-specific iPSC lines were established and characterized as described earlier [9]. Studied cell lines included four HCM cell lines generated from HCM patients carrying either α-tropomyosin (TPM1) or myosin-binding protein C (MYBPC3) mutations, and two LQT1 cell lines generated from patients carrying potassium voltage-gated channel subfamily Q member 1 (KCNQ1) mutations; and four LQT2 cell lines generated from patients carrying mutation in human ether-a-go-go-related gene (hERG). IPSCs were differentiated into spontaneously beating CMs with END2-differentiation method [17] and dissociated into single cell level for Ca^{2+} imaging studies, which was conducted in spontaneously beating Fura-2 AM (Invitrogen, Molecular Probes) or Fluo-4 AM (Life Technologies Ltd) — loaded CMs as described earlier [18]. Ca^{2+} measurements were conducted on an inverted IX70 microscope with a UApo/340 x20 air objective (both Olympus Corporation, Hamburg, Germany) or with Axio Observer.A1 microscope with an Objective Fluar 20x/0.75 M27 (both Carl Zeiss Microscopy GmbH, Göttingen, Germany). Images were taken with an ANDOR iXon 885 CCD camera (Andor Technology, Belfast, Northern Ireland) and synchronized with a Polychrome V light source by a real time DSP control unit or with Lambda DG-4 Plus (Sutter Instrument, California, USA) wavelength switcher and TILLvisION, Live Acquisition (TILL Photonics, Munich, Germany) or ZEN 2 blue edition software (Carl Zeiss Microscopy GmbH, Göttingen, Germany) software. For further Ca^{2+} signal analysis, regions of interest (ROIs) were selected for spontaneously beating cardiomyocytes and background noise was subtracted before further processing. Each Ca^{2+} signal corresponded to a recording from one cardiomyocyte.

3. Peak data extracted from calcium transient signals

The Ca^{2+} signal data measured from cardiomyocytes contains cycles or peaks of different forms, typically quite harmonious in their form and frequency for normally beating iPSC cardiomyocytes, but, on the other hand, deformed and slightly more irregularly occurring peaks for abnormally beating cardiomyocytes. Normality and abnormality of calcium transient signals were determined by a human expert or an algorithm [14,15]. If a calcium transient signal consisted of one or more abnormally developed peaks, such a signal was determined to be abnormal. If a signal included only normally developed peaks, it was determined to be normal. Earlier we categorized them visually to the type of normal and abnormal transient signals, but, as mentioned above, after having discovered that they can be classified together [6,9] we classified them with both types together in the present research. Fig. 1 presents examples from HCMM and HCMT transient signals and Fig. 2 examples from LQT1 and LQT2 signals.

There were 270 HCMM transient signals from two cell lines and 149 HCMT signals from two cell lines, and 90 LQT1 transient signals from two cell lines and 138 LQT2 signals from four cell lines in the data which consequently originated from 10 patients. Roughly a half in both diseases originated from normally beating and the rest from abnormally beating cardiomyocytes. The sampling frequencies were approximately 23 Hz for HCMM, 23 Hz for 95 HCMT signals and 14 Hz for 54 HCMT signals, 8 Hz for LQT1 and 33 Hz for LQT2. The sampling frequency of Ca^{2+} transient measurements has increased in time with the updated measuring system, which explains its variation in the gained data.

For the recognition of peaks from a calcium transient signal, the first derivative approximation was computed throughout the entire signal by estimating first derivative as slope values of linear regression along the signal by using short signal segments each of which included a few samples (amplitude values). The beginning of a peak was recognized when first derivative values increased above a small positive threshold, and then shortly later when first derivative values became less through the left side of the peak and back close to zero, the maximum of the peak was recognized. Thereafter, first derivative values became negative while these were computed from the right side of the peak. Finally, at the end of the peak first derivative values again approached to zero and the end of the peak was reached. Nevertheless, very small peaks were left out as possible noise. To perform this, at first the distribution of amplitude (signal sample) values were computed and the mean of 15% of their largest values was calculated which was used as a rough estimate for peak maxima in the signal. The difference of this estimate and signal

minimum amplitude was compared with a peak candidate. If the peak candidate had a smaller mean of its left and right side amplitudes than approximately 8% from the afore-mentioned estimate of the peak maxima of the entire signal, the peak candidate was left out. This straightforward peak recognition procedure was described more in detail in our previous research [14-16].

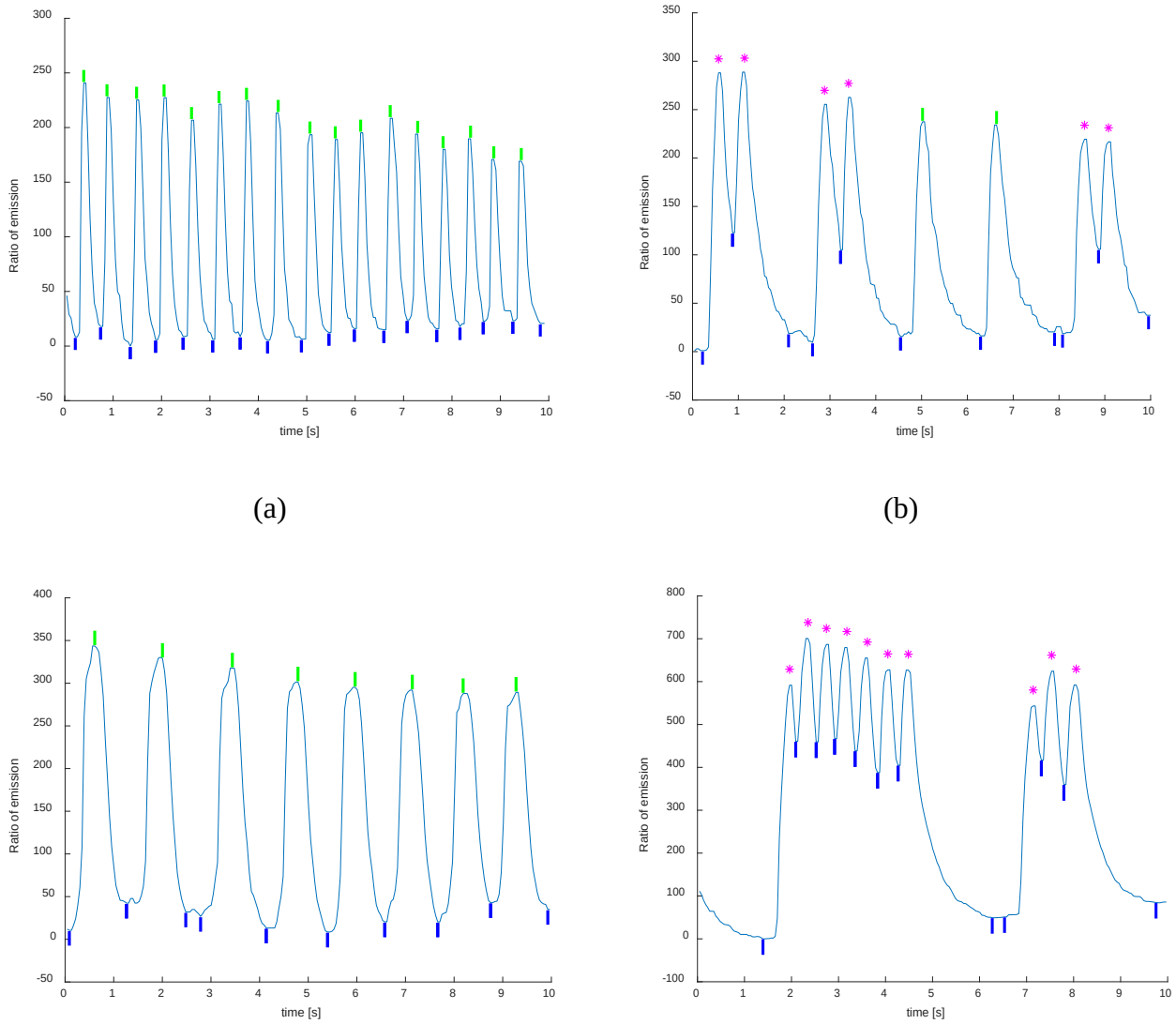


Fig. 1. Segments of 10 s (a) from a normal HCMM transient signal, (b) abnormal (because of three “double peaks”) HCMM signal, (c) normal HCMT signal and (d) abnormal (“multiple peaks”) HCMT signal. The recognition procedure recognized peaks to be either normal marked with a green bar or abnormal marked with a magenta star.

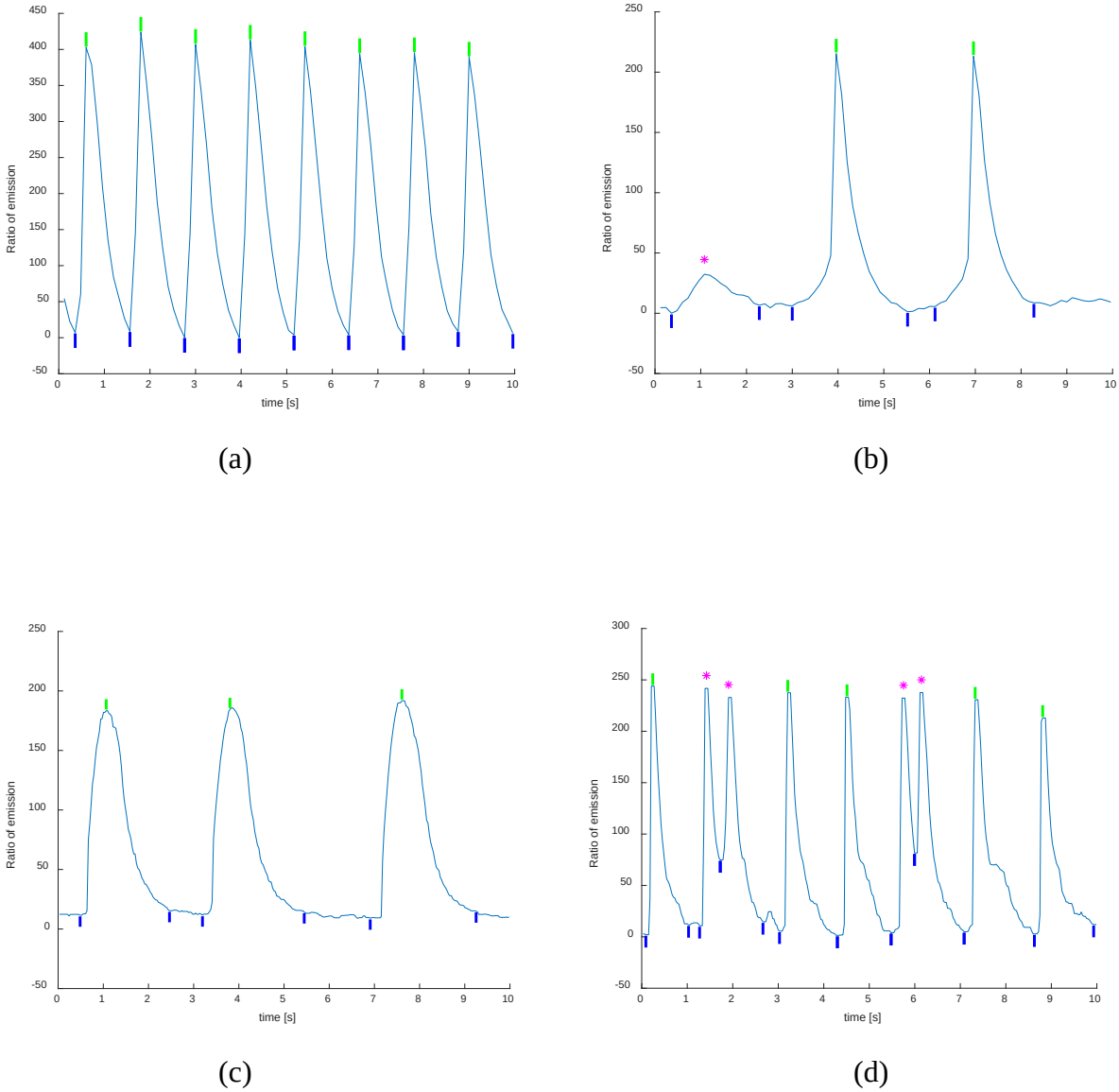
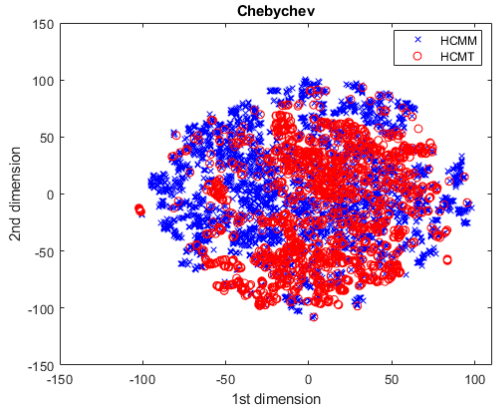
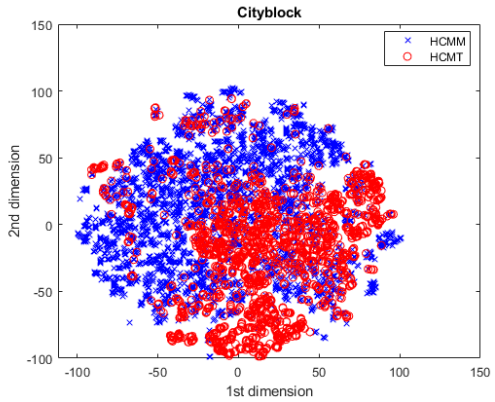


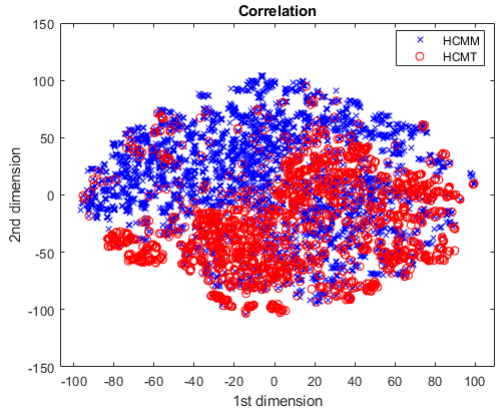
Fig. 2. Segments of 10 s (a) from a normal LQT1 transient signal, (b) abnormal (because of a very low peak) LQT1 signal, (c) normal LQT2 signal and (d) abnormal (two “double peaks”) LQT2 signal. The recognition procedure recognized peaks either normal marked with a green bar or abnormal marked with a magenta star.

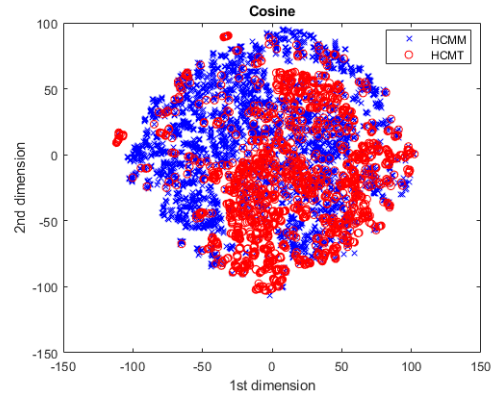
After having recognized the peaks of a transient signal, peak attributes were computed. Earlier [14] we began with seven attributes: left side and right side peak amplitudes, left and right side durations, maximum of the first derivative values from the left and its absolute minimum form the right side, and peak interval computed from the maximum of the preceding peak to the maximum of the current peak. If the current peak was the first one, the interval was computed from the signal beginning. The peak interval attribute characterizes regularity or irregularity of cardiomyocyte cycling. Later [9,16], we took five additional attributes: the maximum and absolute minimum of the second derivative values from the right peak side (left side not used typically being short), surface area given by the peak curve and the line from the peak beginning to its end, duration from the peak beginning to the location of the first derivative maximum within the peak left side, and duration from the location of the first derivative absolute minimum within the peak right side to the peak maximum. As new in the current research we employed two following attributes. Along the peak left and right amplitudes their amplitude halves were computed, then the mean of those two halves and finally the intersection points of the mean and the peak curve. The first new attribute called mean peak duration was equal to the length or actually duration from the left intersection point to the right one. The second new attribute was formed by computing the length of the peak curve from the peak beginning to its end by adding, one by one, consecutive Euclidean distances from the preceding sample (amplitude value) to the current.

The numbers of peaks found from HCMM and HCMT transient signals were 4413 and 2136, respectively. Further, 1635 and 3870 peaks were recognized from LQT1 and LQT2 signals. The peak data generated can be visualized by means of t- Distributed Stochastic Neighbor Embedding [32,33] algorithm implemented in Matlab, which produced data scatter illustrations in Fig. 3 for HCMM and HCMT and Fig. 4 for LQT1 and LQT2 when eight different distance measures were applied. When two mutations are separate from each other, these predict good opportunities for their classification, i.e., separation of both pairs.


Chebychev
HCMM
HCMT
2nd dimension
1st dimension
150
100
50
0
-50
-100
-150
-150
-100
-50
0
50
100


Cityblock
HCMM
HCMT
2nd dimension
1st dimension
150
100
50
0
-50
-100
-100
-50
0
50
100
150


Correlation
HCMM
HCMT
2nd dimension
1st dimension
150
100
50
0
-50
-100
-150
-100
-80
-60
-40
-20
0
20
40
60
80
100


Cosine
HCMM
HCMT
2nd dimension
1st dimension
100
50
0
-50
-100
-150
-150
-100
-50
0
50
100
150

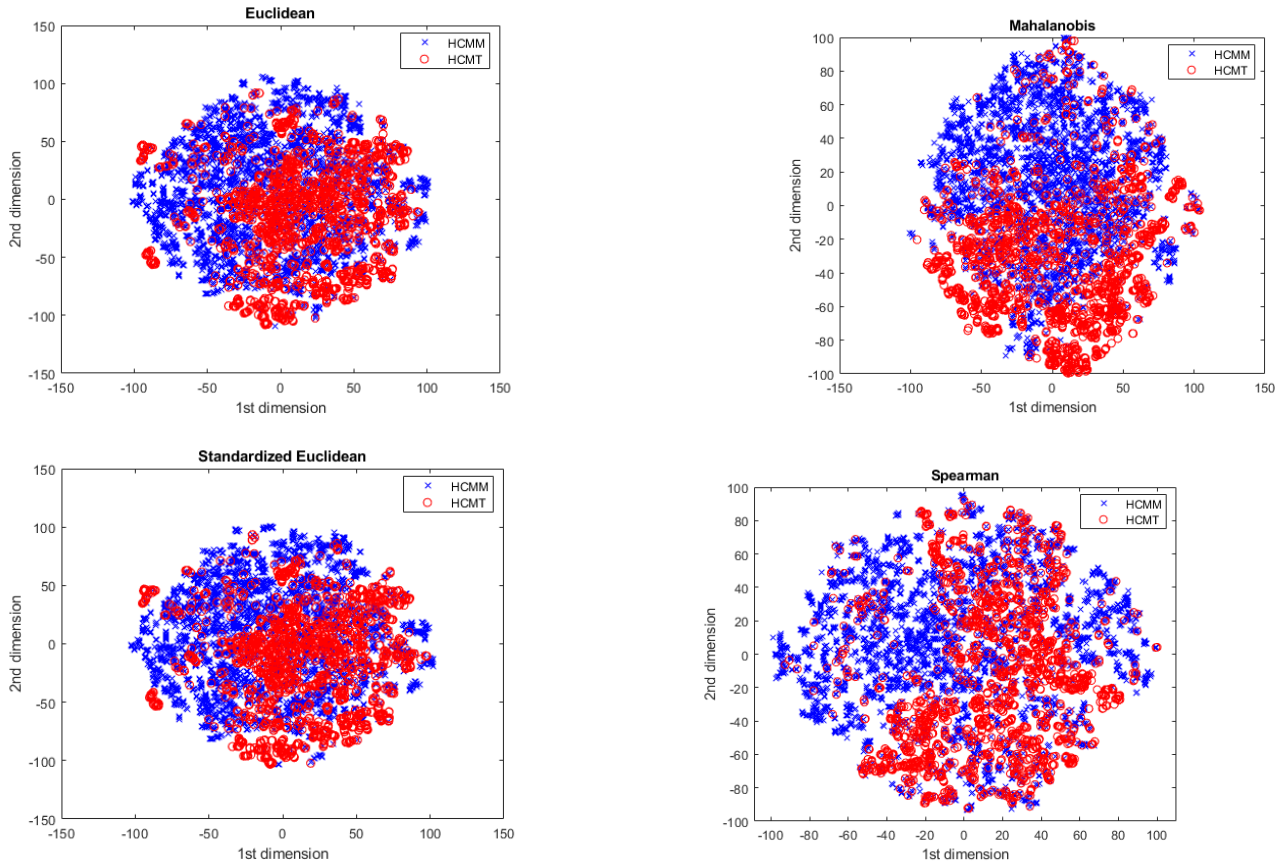
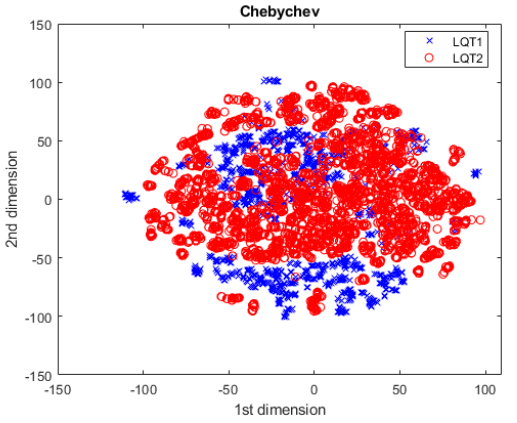
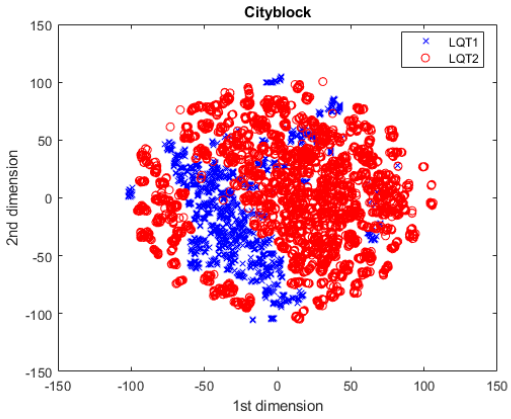
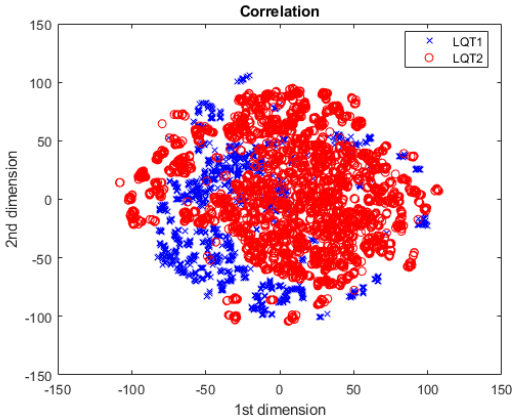
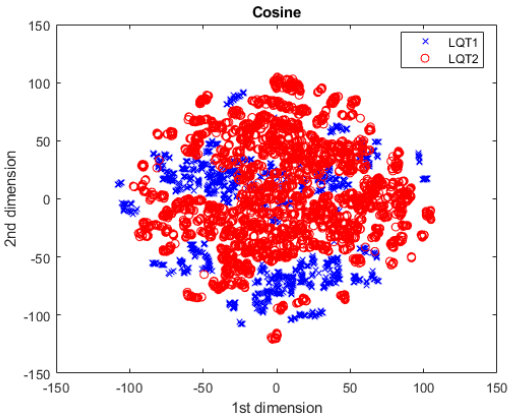


Fig. 3 HCMM and HCMT peak data transformed from a 14-dimensional attribute space to a planar visualization using t-SNE algorithm. In all subfigures perplexity value of 5 was used and the dataset was z-score standardized to have mean of zero and unit variance before visualization.


Chebychev
LQT1
LQT2
2nd dimension
1st dimension
150
100
50
0
-50
-100
-150
-150
-100
-50
0
50
100


Cityblock
LQT1
LQT2
2nd dimension
1st dimension
150
100
50
0
-50
-100
-150
-150
-100
-50
0
50
100
150


Correlation
LQT1
LQT2
2nd dimension
1st dimension
150
100
50
0
-50
-100
-150
-150
-100
-50
0
50
100
150


Cosine
LQT1
LQT2
2nd dimension
1st dimension
150
100
50
0
-50
-100
-150
-150
-100
-50
0
50
100
150

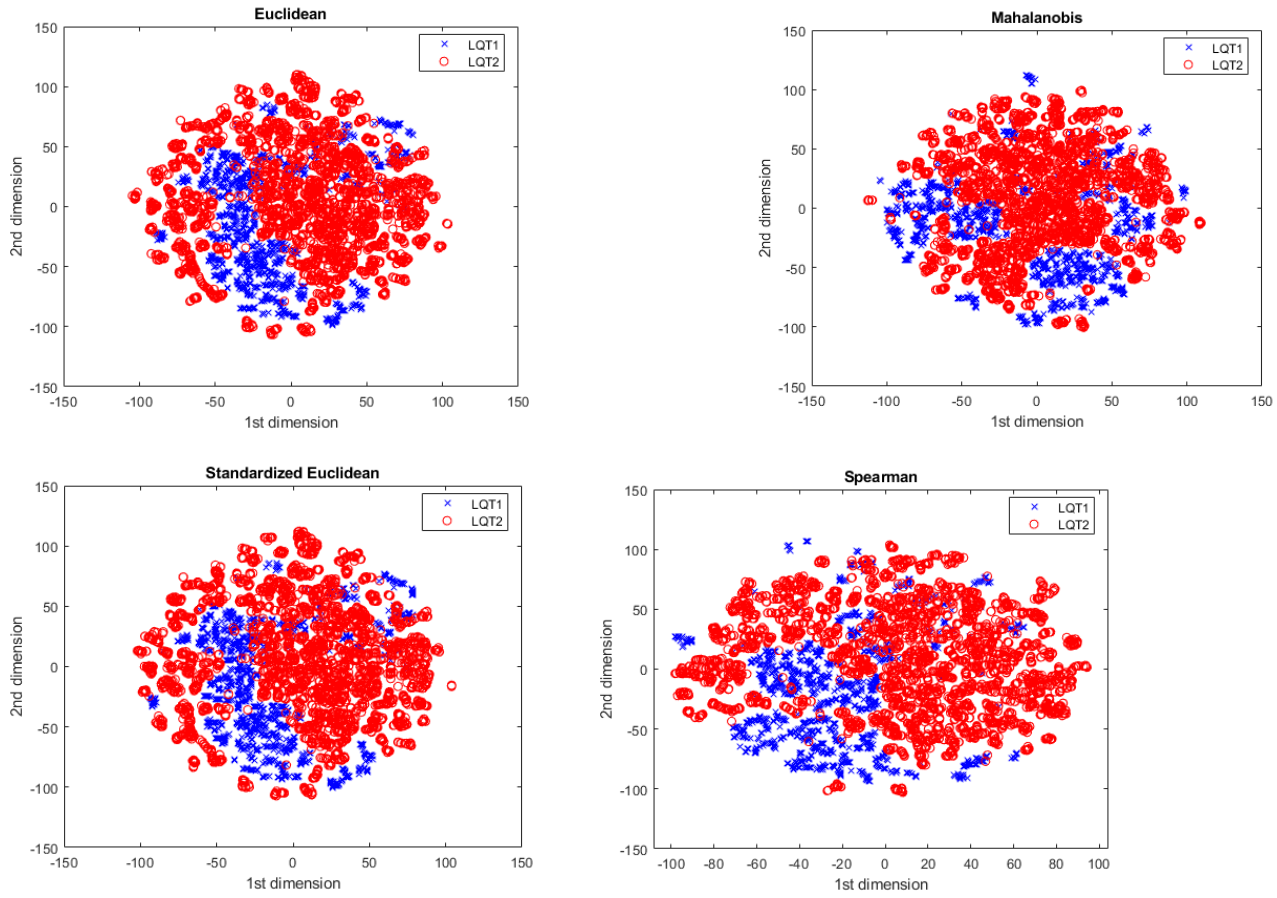


Fig. 4. LQT1 and LQT2 peak data transformed from a 14-dimensional attribute space to a planar visualization using t-SNE algorithm. In all subfigures perplexity value of 5 was used and the dataset was z-score standardized to a mean of zero and unit variance before visualization.

The means and standard deviations of 14 attributes were computed for two mutation pairs. These are given in Table 1. There are clear differences at least for the means of one pair from almost all attributes. This was promising for the separation of the mutations of two diseases.

Table 1. Means and standard deviations of 14 peaks attributes calculated. Six of them are dependent on time in seconds [s]. The rest are dependent on amplitude (without unit being ratios) only or both.

Peak attributes	HCMM	HCMT	LQT1	LQT2
Left amplitude	197.9±92.4	198.7±134.8	170.2±78.6	138.3±90.1
Right amplitude	200.3±93.7	203.0±137.8	172.0±79.5	138.5±97.9
Left duration [s]	0.27±0.15	0.38±0.19	0.33±0.18	0.33±0.24
Right duration [s]	0.48±0.26	0.51±0.33	0.68±0.40	0.76±0.57
First derivative maximum of left side	1952±888	1468±1041	818±472	1596±957
First derivative absolute minimum of right side	1030±444	903±479	509±259	555±296
Second derivative maximum of right side	6002±3190	4363±3050	1615±1324	3373±2722
Second derivative absolute minimum of right side	3433±3099	3405±3645	1208±1432	2814±2994
Peak surface area	50.85±43.48	61.81±69.21	57.69±42.12	59.31±81.41
Peak interval [s]	0.80±0.50	1.02±0.71	1.17±0.92	1.45±1.88
Time difference from the first derivative maximum location of left side to peak beginning [s]	0.19±0.12	0.29±0.18	0.22±0.15	0.18±0.15
Time difference from the first derivative absolute minimum location of right side to peak maximum [s]	0.12±0.07	0.11±0.05	0.15±0.08	0.14±0.17
Mean peak duration [s]	0.26±0.11	0.28±0.10	0.40±0.15	0.38±0.27
Length of peak curve	415.7±190.1	421.2±279.0	343.6±156.9	298.1±186.6

4. Separation of mutations by means of machine learning

Separation of two mutations of both genetic cardiac diseases was computed by applying several machine learning methods in order to study the best classification methods for this purpose. For this study the same classification methods were used as in [16] and the methods were *k*-NN [19,20], linear discriminant analysis (LDA) [21], Mahalanobis discriminant analysis [22], quadratic discriminant analysis [23], classification and regression trees (CART) [24], multinomial logistic regression in two-classes case [25] that basically is logistic regression, naïve Bayes with and without kernel density estimation (KDE) [26,27], random forests [28,29], and least-squares support vector machines (LS-SVM) [30,31]. Before classification, the dataset was z-score standardized so that each attribute has mean of zero and unit variance. This procedure ensures that each attribute is equally important in a dataset.

Parameters play a key role in machine learning. In this study we investigated extensively the effect of parameters. With *k*-NN we varied the *k* value, distance measure and weighting scheme. The *k* values tested belong to set {1,3,5,…,37}, whereas distance measures were {Chebychev, cityblock, correlation, cosine, Euclidean, Mahalanobis, standardized Euclidean, Spearman} and weighting schemes {equal weights, inverse weighting, squared inverse weighting} respectively. In the case of naïve Bayes we tried four kernels in KDE and these were {Gaussian kernel, Epanechnikov kernel, box kernel, triangle kernel}. For random forest classifiers we tested the number of trees from 1 to 100 with step size of 1. The case where only one tree is in a forest differs from basic decision tree produced by CART, because in random forest classifiers the attributes that are used to construct a tree are selected randomly, whereas CART uses all available attributes. The performance of LS-SVM is highly dependent on the selected hyperparameter values. In this study we applied four kernels (linear, quadratic, cubic, RBF) and the parameter value space for the C (boxconstraint) and σ was $\{2^{-12}, 2^{-11},\ldots, 2^{17},\}$.

The actual classification was performed using leave-one-signal-out (LOSO) approach that is designed for signal classification purposes. The idea behind LOSO is similar to leave-one-out, commonly used in machine learning, but now in each cross-validation round the data from one signal in total is left for test set and the rest of the data forms a training set. Training of a classifier is done with a peak-based data and a trained classifier gives a predicted class label for each peak encountered in a test set. In order to obtain a signal level classification, we compute the mode from the predicted class labels for peaks in a test set (in LOSO a test set contains only the data from one signal). Finally, mode is the predicted class label for a signal. Hence, we can compare the ground truth signal class label to predicted signal level

class label and we can evaluate the performance measures. Mode can be unambiguously determined and, thus, we need to prepare a procedure for possible ties also. If a tie occurs, we evaluate the class sizes in a training set and their corresponding proportions with respect to the size of a training set. We divide the interval [0,1] with the obtained proportions into two subintervals where each one of the subintervals is for one class. Then, we produce a random number from uniform distribution *U(0,1)* and find the subinterval where the generated random number belongs to. The subinterval obtained determines the final class label for the signal, if a tie has been encountered. If a classification method included parameter values, LOSO procedure was repeated with all parameter values tested and the best parameter value was determined based on the highest accuracy that is defined in detail in the next section. Overall, the classification procedure used in this study followed the same pattern as what was described in [16] and from [16] more details related to the classification procedure can be seen.

5. Results

The classification procedure implemented is explained in the following and results gained are given in Tables 2 and 3, where the following classification results were computed with each method: true positive rates for HCMM and HCMT, classification accuracy, F1 score and Matthew correlation coefficient. The corresponding results are shown for LQT1 and LQT2 in Tables 4 and 5. True positive rates *TPR* were calculated by counting correctly classified *TP* (true positive, one mutation) and *TN* (other mutation) for both mutations of each disease. Classification accuracy was obtained by using those and the total of transient signals per disease containing also false positive *FP* and negative *FN*, all signals as *N*. In addition, *F1* and Matthew correlation coefficient *MCC* were computed. The last one is appropriate for biased class (mutation here) distributions.

$$TPR = \frac{TP}{TP + FN} 100\%$$

$$A = \frac{TP + TN}{TP + TN + FP + FN} 100\% = \frac{TP + TN}{N} 100\%$$

$$F1 = \frac{2TP}{2TP + FP + FN} 100\%$$

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} 100\%$$

The best accuracy results in Table 2 are 89% and several other are 88%. Thus, the nearest neighbor searching with appropriate parameter values were the best, while the best in Table 3 was also 89% generated by random forests. In Table 4 nearest neighbor searching and in Table 5 several methods obtained exceptionally high results of 100%, which show how LQT1 and LQT2 differ very clearly from each other. Fig. 3 illustrates HCMM and HCMT peak data and indicates slightly more overlapping than Fig. 4 of LQT1 and LQT2 peak data in which these two mutations are located almost fully apart from each other.

Table 2. HCMM vs. HCMT: Results of *k* nearest neighbor (*k*NN) searching by different distance measures with the best *k* value (number of nearest neighbors). The best results are marked in Bold.

Classification method	True positive rate of HCMM %	True positive rate of HCMT %	Accuracy %	F1 score %	Matthew correlation coefficient %
Chebychev metric and equal weighting, *k*=1	93	72	85	89	67
Chebychev metric and inverse weighting, *k*=1	93	72	85	89	67
Chebychev metric and squared inverse weighting, *k*=1	93	72	85	89	67
cityblock metric and equal weighting, *k*=1	94	79	**89**	**91**	**75**
cityblock metric and inverse weighting, *k*=1	94	79	**89**	**91**	**75**
cityblock metric and squared inverse weighting, *k*=1	94	79	**89**	**91**	**75**
correlation measure and equal weighting, *k*=1	92	76	86	90	69
correlation measure and inverse weighting, *k*=5	90	83	87	90	72
correlation measure and squared inverse weighting, *k*=25	90	83	88	90	73
cosine measure and equal weighting, *k*=1	92	75	86	89	69
cosine measure and inverse weighting, *k*=11	90	79	86	89	70
cosine measure and squared inverse weighting, *k*=5	93	76	87	90	70
Euclidean metric and equal weighting, *k*=1	93	79	88	91	74
Euclidean metric and inverse weighting, *k*=1	93	79	88	91	74
Euclidean metric and squared inverse weighting, *k*=1	93	79	88	91	74
Mahalanobis metric and equal weighting, *k*=3	95	76	88	91	74
Mahalanobis metric and inverse weighting, *k*=3	95	76	88	91	74
Mahalanobis metric and squared inverse weighting, *k*=3	95	75	88	91	73
standardized Euclidean metric and equal weighting. *k*=1	93	79	88	91	74

standardized Euclidean metric and inverse weighting, k=1	93	79	88	91	74
standardized Euclidean metric and squared inverse weighting, k=1	93	79	88	91	74
Spearman measure and equal weighting, k=1	92	75	86	90	69
Spearman measure and inverse weighting, k=1	92	75	86	90	69
Spearman measure and squared inverse weighting, k=1	92	75	86	90	69

Table 3. HCMM vs. HCMT: Results of discriminant analysis, Bayesian methods, decision trees, random forests, and support vector machines. The best results are marked in Bold. LS-SVM equals least squared support vector machine.

Classification method	True positive rate of HCMM %	True positive rate of HCMT %	Accuracy %	F1 score %	Matthew correlation coefficient %
Linear discriminant analysis	87	69	80	85	57
Mahalanobis discriminant analysis	62	95	74	75	55
Quadratic discriminant analysis	81	71	78	82	52
Decision trees	94	71	85	89	67
Multinomial logistic regression	87	72	81	86	59
Naïve Bayes with normal distribution	74	60	69	75	33
Naïve Bayes with normal kernel	76	74	75	80	48
Naïve Bayes with box kernel	80	66	75	80	46
Naïve Bayes with Epanechnikov kernel	72	73	72	77	43
Naïve Bayes with triangle kernel	77	72	75	80	47
Random forest, number of trees 49	95	78	**89**	**92**	**75**
LS-SVM with linear kernel, $C=2^{16}$	76	87	80	83	60
LS-SVM with quadratic kernel, $C=2^{4}$	82	83	82	86	63
LS-SVM with cubic kernel, $C=2^{-7}$	86	83	85	88	67
LS-SVM with RBF kernel, C=32, σ=1	94	78	**88**	**91**	**74**

Table 4. LQT1 vs. LQT2: Results of *k* nearest neighbor (*k*NN) searching by different distance measures with the best *k* value (number of nearest neighbors). The best results are marked in Bold.

Classification method	True positive rate of LQT1 %	True positive rate of LQT2 %	Accuracy %	F1 score %	Matthew correlation coefficient %
Chebychev metric and equal weighting, *k*=1	91	99	96	95	92
Chebychev metric and inverse weighting, *k*=1	91	99	96	95	92
Chebychev metric and squared inverse weighting, *k*=1	91	99	96	95	92
cityblock metric and equal weighting, *k*=1	96	100	98	98	96
cityblock metric and inverse weighting, *k*=1	96	100	98	98	96
cityblock metric and squared inverse weighting, *k*=1	96	100	98	98	96
correlation measure and equal weighting, *k*=7	98	98	98	97	95
correlation measure and inverse weighting, *k*=7	98	97	97	97	95
correlation measure and squared inverse weighting, *k*=9	98	96	97	96	94
cosine measure and equal weighting, *k*=1	97	99	98	98	96
cosine measure and inverse weighting, *k*=1	97	99	98	98	96
cosine measure and squared inverse weighting, *k*=1	97	99	98	98	96
Euclidean metric and equal weighting, *k*=1	93	99	97	96	94
Euclidean metric and inverse weighting, *k*=1	93	99	97	96	94
Euclidean metric and squared inverse weighting, *k*=1	93	99	97	96	94
Mahalanobis metric and equal weighting, *k*=13	100	100	**100**	**100**	**100**
Mahalanobis metric and inverse weighting, *k*=13	100	100	**100**	**100**	**100**
Mahalanobis metric and squared inverse weighting, *k*=1	100	99	**100**	**100**	**100**
standardized Euclidean metric and equal weighting. *k*=1	93	99	97	96	94
standardized Euclidean metric and inverse weighting, *k*=1	93	99	97	96	94
standardized Euclidean metric and squared inverse weighting, *k*=1	93	99	97	96	94
Spearman measure and equal weighting, *k*=3	100	98	99	98	97
Spearman measure and inverse weighting, *k*=3	100	98	99	98	97
Spearman measure and squared inverse weighting, *k*=3	100	98	99	98	97

Table 5. LQT1 vs. LQT2: Results of discriminant analysis, Bayesian methods, decision trees, random forests, and support vector machines. The best results are marked in Bold. LS-SVM equals least squared support vector machine.

Classification method	True positive rate of LQT1 %	True positive rate of LQT2 %	Accuracy %	F1 score %	Matthew correlation coefficient %
Linear discriminant analysis	92	97	95	94	90
Mahalanobis discriminant analysis	70	100	88	82	77
Quadratic discriminant analysis	91	96	94	92	87
Decision trees	100	100	**100**	**100**	**100**
Multinomial logistic regression	98	100	99	99	98
Naïve Bayes with normal distribution	74	77	75	70	50
Naïve Bayes with normal kernel	82	99	92	89	84
Naïve Bayes with box kernel	78	99	90	86	80
Naïve Bayes with Epanechnikov kernel	79	99	91	87	81
Naïve Bayes with triangle kernel	83	99	93	90	85
Random forest, number of trees 4	100	100	**100**	**100**	**100**
LS-SVM with linear kernel, $C=2^{-1}$	100	97	98	98	96
LS-SVM with quadratic kernel, $C=2^{-7}$	100	100	**100**	**100**	**100**
LS-SVM with cubic kernel, $C=2^{-8}$	100	99	100	99	99
LS-SVM with RBF kernel, $C=2^{3}$, $\sigma=2^{2}$	100	100	**100**	**100**	**100**

6. Discussion

Our results showed that with computational machine learning method both HCMM and HCMT mutations as well as LQT1 and LQT2 disease types could be separated from each other by Ca^{2+} transient signals with high accuracy. This replenishes our previous findings and shows the possibility to discriminate genetic cardiac diseases and even different mutations by Ca^{2+} transient profiles recorded from iPSC-CMs with machine learning classification methods. Computational machine learning method

of iPSC-CM Ca^{2+} transient profiles could be an automated, high throughput method to help diagnostics in the future.

According to true positive rates we see how HCMT was more difficult to be separated than HCMM. Sometimes, HCMT signals were classified incorrectly to HCMM class, but more infrequently on the contrary. A probable cause is that HCMM was predominant containing more signals, 270 vs. only 149 in HCMT. A quite similar situation exists between the true positive rate results of LQT1 and LQT2, where the latter is predominant, 90 vs. 138 transient signals. However, this is not so prominent for LQT1 and LQT2, when their true positive rates are very close or equal to 100%.

The sampling frequency f of 8 Hz for LQT1 corresponds to around 0.12 s as time interval T when $f=1/T$. Further, 33 Hz of LQT2 corresponds to around 0.03 s. The different time intervals do not base an ideal situation. Six of the attributes in Table 1 depend directly on time, two depend on amplitude and the other six depend on both time and amplitude. The difference of two time intervals of 0.12-0.03=0.09 s can be seen as a maximal “error” of the lower sampling frequency compared to the higher frequency. Because sampling a signal is of quantization, various possible locations of a peak beginning, maximum or end are equally probable within the difference of 0.09 s, i.e., those possible locations are uniformly distributed in principle. Thus, it is a half of 0.09 s on the average. On the other hand, such a difference could affect both the beginning and end locations of each directly time dependent attribute. Their common effect would then be 2 times 0.09/2 = 0.09 s. Five of six attributes of LQT1 and LQT2 depending directly on time contain quite similar means close to each other, when their differences are less than 0.09 s. One attribute called peak interval has a greater difference between LQT1 and LQT2. Thus, other than time dependent attributes are the most important for the separation of lQT1 and LQT2.

For those attributes depending on amplitude or both amplitude and time, differences between LQT1 and LQT2 are considerable for most attributes. Since classification results were exceptionally good for LQT1 and LQT2, the reason is that their attribute values differ greatly from each other for other attributes than the directly time dependent. For HCMM and HCMT, such a marginal inaccuracy because of sampling frequencies was non-existent, because the higher sampling frequencies used for HCMT was the same to that of HCMM and also contained the majority from among HCMT signals.

7. Conclusion

All in all, in the current data HCMM and HCMT mutations could be separated very properly from each other up to the accuracy of 89% and LQT1 and LQT2 mutations exceptionally well up to the accuracy of 100%. These very good results denote a great potential of machine learning separation of two types of both diseases. These results reinforce our previous findings and enlarges the possibilities of our method to distinguish and identify wide spectrum of different genetic cardiac diseases. This machine learning classification method could be exploited to diagnose genetic cardiac disease and could even predict the type of mutation based on only Ca^{2+} transient signals measured from iPSC-CMs. In the future, this method could be used to diagnose and stratify patients into different treatment groups as well as optimize patient specific drug therapy and treatment efficiency. We will collect new data with regard in the current two cardiac diseases and also some other not yet studied this way at all. Furthermore, we will tailor the applied machine learning methods and others not yet applied and develop them to various calcium transient signal classification tasks of iPSC-CMs.

Summary table

- On the basis of our earlier research we knew that it is possible to separate such genetic cardiac diseases as HCM and LQT from each other by using iPSC calcium transient signals as data for machine learning methods.
- The present study shows that it is also possible to efficiently separate LQT1 from LQT2 originating from different mutations and correspondingly HCMM from HCMT by using iPSC calcium transient signals and machine learning.

Acknowledgements

The work of the second author was supported by The Finnish Foundation for Cardiovascular Research and The Maud Kuistila Memorial Foundation.

Author contributions

KP and KA-S designed the cell experiments. KP made the cell experiments and analyzed the raw cell calcium data. MJ programmed and ran the signal analysis phase and computed the peak variable values. HJ executed the classifications by means of machine learning. KA-S contributed reagents, materials and analysis tools. All authors wrote the manuscript.

Declarations of interest: none

References

[1] A. Moretti, M. Bellin, A. Welling, C.B. Jung, J.T. Lam, L. Bott-Flugel, T. Dorn, A. Goedel, C. Hohnke, F. Hofmann, M. Seyfarth, D. Sinnecker, A. Schomig, K.L., Laugwitz, Patient-specific induced pluripotent stem-cell models for long-QT syndrome, N. Engl. J. Med. 363 (2010) 1397-409.

[2] E. Matsa, D. Rajamohan, E. Dick, L. Young, I. Mellor, A. Staniforth, C. Denning, Drug evaluation in cardiomyocytes derived from human induced pluripotent stem cells carrying a long QT syndrome type 2 mutation, Eur. Heart J. 32 (2011) 952-962.

[3] A.L. Lahti, V.J. Kujala, H. Chapman, A.P. Koivisto, M. Pekkanen-Mattila, E. Kerkela, J. Hyttinen, K. Kontula, H. Swan, B.R. Conklin, S. Yamanaka, O. Silvennoinen, K. Aalto-Setala, Model for long QT syndrome type 2 using human iPS cells demonstrates arrhythmogenic characteristics in cell culture, Dis. Model. Mech. 5 (2012) 220-230.

[4] J. Kuusela, J. Kim, E. Räsänen, K. Aalto-Setälä, The Effects of Pharmacological Compounds on Beat Rate Variations in Human Long QT-Syndrome Cardiomyocytes, Stem Cell Rev. 12(6) (2016) 698-707.

[5] M. Ojala, C. Prajapati, R.P. Pölönen, K. Rajala, M. Pekkanen-Mattila, J. Rasku, K. Larsson, K. Aalto-Setälä, Mutation-specific phenotypes in hiPSC-derived cardiomyocytes carrying either myosin-binding protein C or α-tropomyosin mutation for hypertrophic cardiomyopathy, Stem Cells Int. 2016. https://www.hindawi.com/journals/sci/2016/1684792/

[6] L. Han, Y. Li, J. Tchao, A.D. Kaplan, B. Lin, Y. Li, J. Mich-Basso, A. Lis, N. Hassan, B. London et al., Study familial hypertrophic cardiomyopathy using patient-specific induced pluripotent stem cells, Cardiovasc. Res. 104 (2014) 258-269. 10.1093/cvr/cvu205

[7] F. Lan, A.S. Lee, P. Liang, V. Sanchez-Freire, P.K. Nguyen, L. Wang, L. Han, M. Yen, Y. Wang, N. Sun et al., Abnormal calcium handling properties underlie familial hypertrophic cardiomyopathy pathology in patient-specific induced pluripotent stem cells, Cell Stem Cell 12 (2013) 101-113. 10.1016/j.stem.2012.10.010

[8] K. Penttinen, H. Swan, S. Vanninen, J. Paavola, A.M. Lahtinen, K. Kontula, K. Aalto-Setälä, Antiarrhythmic effects of Dantrolene in patients with catecholaminergic polymorphic ventricular tachycardia and replication of the responses using iPSC models, Plos One 10(7) (2015).

[9] M. Juhola, H. Joutsijoki, K. Penttinen, K. Aalto-Setälä, Detection of genetic cardiac diseases by Ca^{2+} transient profiles using machine learning methods, Sci. Rep. 8 (2018) Article number 9355. doi:10.1038/s41598-018-27695-5

[10] B.J. Maron, S.R. Ommen, C. Semsarian, P. Spirito, I. Olivotto, and M.S. Maron, Hypertrophic cardiomyopathy: present and future, with translation into contemporary cardiovascular medicine, J. American College of Cardiology, 64(1) (2014) 89–99.

[11] Q. Wang, M.E. Curran, I. Splawski, T.C. Burn, J.M. Millholland, T.J. Van Raay, J. Shen, K.W. Timothy, G.M. Vincent, T. de Jager et al., Positional cloning of a novel potassium channel gene: KVLQT1 mutations cause cardiac arrhythmias, Nat. Genet. 12(1) (1996) 17–23.

[12] M.C. Sanguinetti, C. Jiang, M.E. Curran, M.T. Keating, A mechanistic link between an inherited and an acquired cardiac arrhythmia: HERG encodes the IKr potassium channel, Cell 81(2) (1995) 299–307.

[13] E.K. Lee, D.D. Tran, W. Keung, P. Chan, G. Wong, C.W. Chan, K.D. Costa, R.A. Li, M. Khine, Machine learning of human pluripotent stem cell-derived engineered cardiac tissue contractility for automated drug classification, Stem Cell Rep. 9 (2017) 1560-1572.

[14] M. Juhola, H. Joutsijoki, K. Varpa, J. Saarikoski, J. Rasku, K. Iltanen, J. Laurikkala, H. Hyyrö, J. Ávalos-Salguero, H. Siirtola, K. Penttinen, K. Aalto-Setälä, On computation of calcium cycling anomalies in cardiomyocytes data, 36th Annual Int. Conf. IEEE Eng. Med. Biol. Society, 2014, Chicago, Illinois, USA, 1444-1447.

[15] M. Juhola, K. Penttinen, H. Joutsijoki, K. Varpa, J. Saarikoski, J. Rasku, H. Siirtola, K. Iltanen, J. Laurikkala, H. Hyyrö, J. Hyttinen, K. Aalto-Setälä, Signal analysis and classification methods for calcium transient data of stem cell derived cardiomyocytes, Comp. Biol. Med. 61 (2015) 1-7.

[16] M. Juhola, H. Joutsijoki, K. Penttinen, K. Aalto-Setälä, Machine learning to differentiate diseased cardiomyocytes from healthy control cells, Inf. Med. Unlocked 14 (2019) 15-22.

[17] C. Mummery, D. Ward-van Oostwaard, P. Doevendans, R. Spijker, S. van den Brink, R. Hassink, M. van der Heyden, T. Opthof, M. Pera, A.B. de la Riviere, R. Passier, R. Tertoolen, Differentiation of human embryonic stem cells to cardi-omyocytes: role of coculture with visceral endoderm-like cells, Circulation 107 (2003) 2733-2740.

[18] K. Kujala, J. Paavola, A. Lehti, K. Larsson, M. Pekkarinen-Mattila, M. Viitasalo, A.M. Lahtinen, L. Toivonen, K. Kontula, H. Swan, M. Laine, O. Silvennoinen, K. Aalto-Setälä, Cell model of catecholaminercig polymorphic ventricular tachycardia reveals early and delayed after depolarizations, Plos One 7(9) (2012).

[19] P.T. Noi, M. Kappas, Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using Sentinel-2 imagery, Sensors 18(1) (2018) Article 18.

[20] S.A. Dudani, The distance weighted k-nearest neighbor rule, IEEE Transactions on Systems, Man, and Cybernetics 6(4) (1976) 325-327.

[21] A. Tharwat, T. Gaber, A. Ibrahim, A.E. Hassanien, Linear discriminant analysis: A detailed tutorial, AI Communication 30(2017) 169-190.

[22] G. Bohling, Classical normal-based discriminant analysis, Technical Report EECS 833 (2006) 1-24.

[23] A. Tharwat, Linear vs. quadratic discriminant analysis classifier: a tutorial, International Journal of Applied Pattern Recognition 3(2) (2016) 145-179.

[24] E. Gokgoz, A. Subasi, Comparison of decision tree algorithms for EMG signal classification using DWT, Biomedical Signal Processing and Control 18 (2015) 138-144.

[25] C. Kwak, A. Clayton-Matthews, Multinomial logistic regression, Nursing Research 51(6) (2002) 404-410.

[26] H.J. Escalante, E.F. Morales, E. Sucar, A naïve Bayes baseline for early gesture recognition, Pattern Recognition Letters 73 (2016) 91-99.

[27] J.N.K. Liu, Y.-L. He, X.-Z. Wang, Y.-X. Hu, A comparative study among different kernel functions in flexible naïve Bayesian classification, Proceedings of the 2011 International Conference on Machine Learning and Cybernetics (2011) 638-643.

[28] L. Breiman, Random forests, Machine Learning 45(1) (2001) 5-32.

[29] G. Biau, E. Scornet, A random forest guided tour, TEST 25(2) (2016) 197-227.

[30] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Processing Letters 9(3) (1999) 293-300.

[31] A. Yang, W. Li, X. Yang, Short-term electricity load forecasting based on feature selection and least squares support vector machines, Knowledge-Based Systems 163 (2019) 159-173.

[32] L.J.P. van der Maaten, Accelerating t-SNE using tree-based algorithms, Journal of Machine Learning research 15 (2014) 3221-3245.

[33] L.J.P. van der Maaten, G.E. Hinton, Visualizing high-dimensional data using t-SNE, Journal of Machine Learning research 5 (2008) 2579-2605.