Differentiation of Genetic Cardiac Diseases on the Basis of Artificial Intelligence

M Juhola[1]*, H Joutsijoki[1], K Penttinen[2], K Aalto-Setälä[2,3]

[1]Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland

[2]Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland

[3]Heart Center, Tampere University Hospital, Tampere, Finland

Abstract

It has been previously shown that human cardiac disorders can be modeled with induced pluripotent stem cell differentiated cardiomyocytes (iPSC-CM), which enables to study disease characteristics and pathophysiology in more detail. We have shown that some genetic cardiac diseases can be separated from each other and from healthy controls by applying machine learning methods to calcium transient signals measured from these cells. In this study, separation of four genetic cardiac diseases and controls were studied by applying classification methods such as nearest neighbor searching algorithm, decision trees, least squares support vector machines and random forests to peak data computed from calcium transient signals measured from beating induced pluripotent stem cell-derived (iPSC) cardiomyocytes. The best classification accuracy obtained was 77.8% being very promising. The result strengthens our previous finding that the machine learning method can be exploited to identification of several genetic cardiac diseases, but also to separate mutations in different genes resulting in the same clinical phenotype.

Keywords

Induced pluripotent cardiomyocytes; Genetic cardiac diseases; Calcium transient signals; Peak detection; Machine learning; Classification

Correspondence to:

M Juhola
Faculty of Information Technology and Communication Sciences,
Tampere University, Tampere, Finland
E-mail: martti.juhola@tuni.fi

EJBI 2019; 15(3):43-52
Received: June 27, 2019
Accepted: October 28, 2019
Published: November 04, 2019

1. Introduction

Comprehensive functioning of calcium cycling is crucial for excitation-contraction coupling of cardiomyocytes. Abnormal calcium cycling is linked to arrhythmogenesis, which is associated with cardiac disorders and heart failure. Induced pluripotent stem cell-derived cardiomyocytes (iPSC-CMs) [1] have enabled the study of different genetic cardiac diseases. Cardiac diseases can cause changes and variability in calcium cycling that affect the function and phenotype of CMs and previous studies have shown substantial defects and abnormalities in the calcium cycling of iPSC-CMs, reflecting the cardiac phenotype observed in patients [2,3]. Characterizing these disturbances and abnormalities is vital to improve the studies of disease pathology as well as disease diagnostics and treatment.

As we have shown early, machine learning can be utilized as a comprehensive tool for calcium cycling signal analysis of iPSC-CMs [4-6]. Besides that, so far machine learning has obviously been only infrequently used for data extracted from iPSC-CMs. Nevertheless, it has been used to analyze mechanistic action of drugs in cardiology [7] and electrophysiological effects of chronotropic drugs [8].

Previously, we studied how to efficiently separate abnormally beating iPSC-CMs from those normally beating ones [4]. Thereafter, we continued by developing machine learning methods in order to separate diseased cardiomyocytes from those of healthy controls [5] and those of three different diseases from each other and controls [6]. All recognition tasks were based on calcium transient signals measured from beating iPSC-CMs. Roughly a half of transient signals originating from diseased cardiomyocytes are abnormally beating, but only 10-20% in the case of controls' cardiomyocytes [4,5,6]. We then observed that, after all, signals of abnormally and normally beating cardiomyocytes were not necessary to be divided into different groups or the abnormal to be left out, because differentiation between three diseases and controls was possible and made equally well including signals of both normally and abnormally beating cells than performing differentiation separately for the normal and abnormal. Three diseases were long QT syndrome 1 (LQT1), an electric disorder of the heart that may cause arrhythmias, hypertrophic cardiomyopathy (HCM),a disorder that affects the heart muscle tissue structure leading to arrhythmias and a heart failure, and catecholaminergic polymorphic tachycardia (CPVT), a condition characterized by abnormal heart rhythm caused by increase in heart rate.

In the current study, we extend our research as follows. We increased the numbers of calcium transient signals of CPVT and controls (WT, wildtype). In addition, we increased number of HCM signals by including signals from two different HCM disease mutations, which were HCMT, an α-tropomyosin (TPM1) of the β-myosin heavy chain (MYH7) mutation and HCMM, a myosin-binding protein C (MYBPC3) gene mutation [9]. When differentiation is on the basis of different peak forms (beats of iPSC-CMs) in transient signals of the different diseases and controls, we added two new peak attributes compared to those in our earlier research [5].

2. Method

The study was approved by the Ethics Committee of Pirkanmaa Hospital District subject to culturing and differentiating of human iPSC lines (R08070). Patient-specific iPSC lines were established and characterized as described earlier [6]. Studied cell lines included HCM cell lines generated from two HCMT patients carrying α-tropomyosin (TPM1) and two HCMM patients carrying myosin-binding protein C (MYBPC3) mutations, two LQT1 cell lines generated from patients carrying potassium voltage-gated channel subfamily Q member 1 (KCNQ1) mutations; six CPVT cell lines from patients carrying cardiac ryanodine receptor (RyR2) mutation, and one cell line generated from a healthy control individual. IPSCs were differentiated into spontaneously beating CMs with END2-differentiation method [10] 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 [11]. Ca^{2+} measurements were conducted on an inverted IX70 microscope with a UApo/340 20x air objective (both Olympus Corporation, Hamburg, Germany) or with Axio Observer. A1 microscope with a 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. Totally, there were 90 LQT1 calcium transient signals, 149 HCMT signals, 270 HCMM signals, 233 CPVT signals and 199 WT signals.

2.1 Peak Data Derived On The Basis Of Calcium Transient Signals

We have developed an algorithm [4,5] to identify peaks equal to beats from a calcium transient signal. Originally, a biotechnological expert divided signal types to either normally or abnormally beating cardiomyocytes. In possible future applications, this could be automatized for which we have also developed an algorithm [4]. This would be needed for larger quantities of transient signals than now in research. Agreement of the expert and algorithm was approximately 90% subject these two signal types [4].

In the present data there were 62 abnormal and 28 normal calcium transient signals in LQT1, 100 abnormal and 170 normal signals in HCMM, 119 abnormal and 114 normal signals in CPVT, 31 abnormal and 168 normal signals in controls WT, 65 abnormal and 84 normal signals in HCMT. Note that the normal values were relatively much more frequent in WT than in four diseases. However, as mentioned above when both types could be classified virtually equally well into different diseases or controls [6], we did not any longer specify them, but used them as such.

Figure 1 shows 10 s segments from WT signals and Figure 2 from HCMT signals as examples. In general, peak shapes may vary greatly. Nevertheless, for a considerably smaller set of 527 transient signals of three diseases only and controls [6] we observed that they have consistently quite similar properties, subject to peaks, within transient signal sets of individual diseases and that of controls. Therefore, we saw reasonable to extend the number of available diseases and also number of data for most of them, now altogether 941 signals.

In order to identify calcium transient peaks as exactly as possible, the locations of their beginning, maximum and end had to be found. Approximation of the first derivative along a signal was used for this task. When its value was close to zero and rapidly increased to positive values, a peak beginning was identified. Then it decreased back close to zero for a peak maximum and then decreased to negative values, but then again close to zero where a peak end was met. To search for these extreme values, appropriate, small threshold values were found experimentally [4]. Very small peaks with average amplitudes of the left and right peak sides less than approximately 8% of an estimate of large peaks in a signal were not accepted as actual peaks, but rather noise and left out. A large peak estimate was computed as the difference of means of 15% of the greatest values and 15% of least values in the amplitude (sample) distribution of a signal.

After the identification of all acceptable peaks in a signal, attributes of every peak were computed. In our recent research, we used 12 attributes [5]: amplitudes of peak left and right sides, durations [s] of the peak left and right side, the maximum of the left side and absolute minimum of the right side from the approximated first derivative, maximum and absolute minimum of the second derivative of the right side, peak surface area between the peak curve and the line from the peak beginning to the end, duration [s] from the peak maximum to the preceding peak maximum or the beginning of the signal (if the first peak), duration [s] from the peak beginning to the location of the

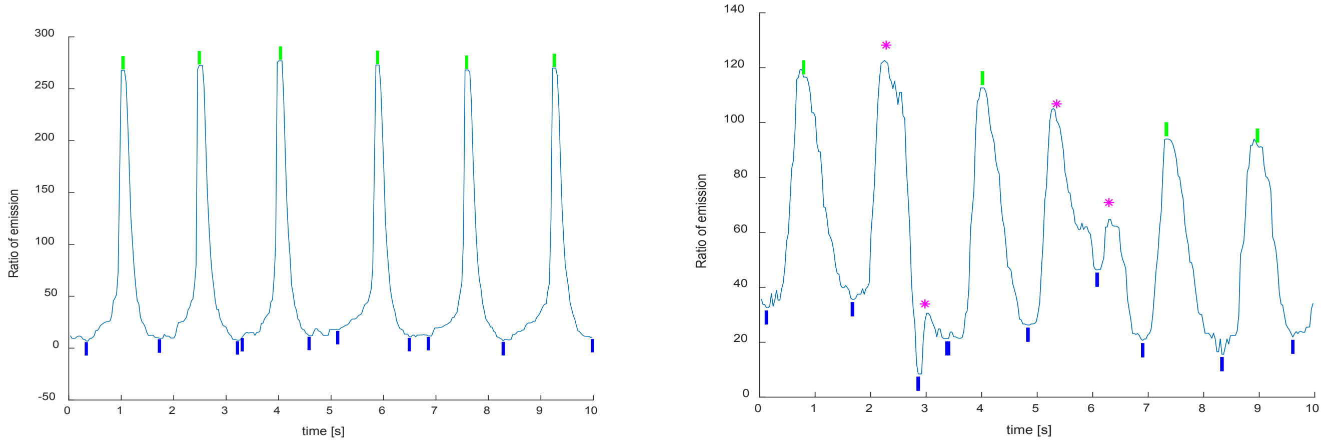

Figure 1: Control WT: (a) A segment of 10 s from a calcium transient signal of a normally beating cardiomyocyte, since all peaks with the green bar of the maximum were recognized to be normal. (b) Correspondingly from an abnormally beating cardiomyocyte, since four peaks with the magenta asterisk were recognized to be abnormal because of different amplitude sizes of left and right peak sides or an entire peak of a considerably less amplitude than the peaks.

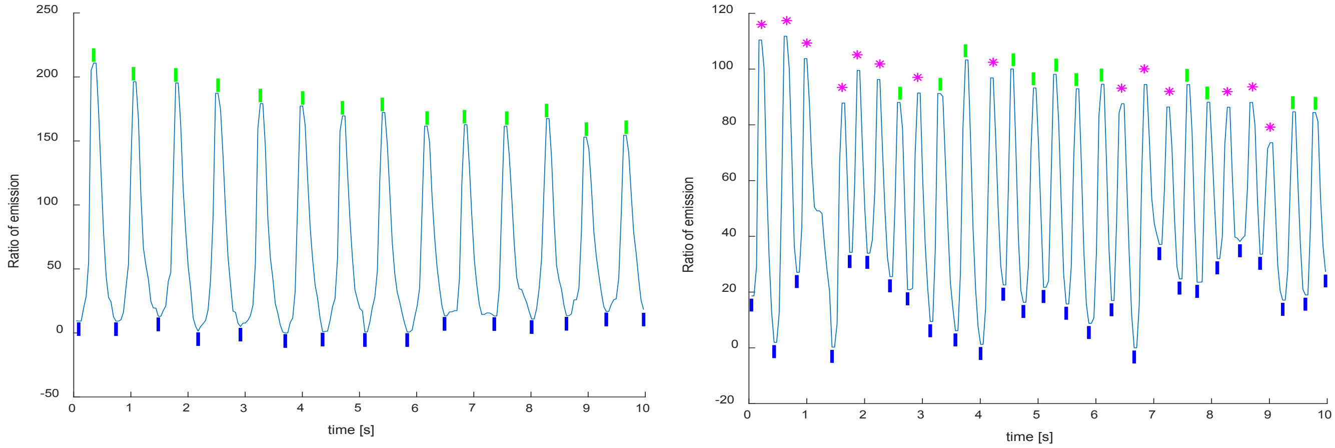

Figure 2: Disease HCMT: (a) A segment of 10 s from a calcium transient signal of a normally beating cardiomyocyte, where all peaks with the green bar of the maximum were recognized to be normal. (b) Correspondingly from an abnormally beating cardiomyocyte, where most peaks with the magenta asterisk were recognized to be abnormal because of different amplitude sizes of left and right peak sides or an entire peak of a considerably less amplitude than the peaks on average.

maximum of the first derivative of the left side, and duration [s] from the peak maximum to the location of the absolute first derivative minimum of the right side. At the present research we still added a new attribute as follows. First, the averages of peak left and right side amplitudes were computed and then the average of these average side amplitudes. Nonetheless, if one of the sides was so small (low) that it was less than a half of the other, the average of the less one was not used, but instead its entire amplitude value from the beginning or end (depending on the side) of the peak to the peak maximum. Finally, the duration [s] called peak average width between the location of the (estimated) average of the left side and that of the right side was computed to be the 13th attribute.

The minimum, mean and maximum lengths of all 941 signals were 7.7, 17.7 and 46.5 s. The minimum, mean and maximum of peak numbers per signal were 1, 13.6 and 61. Figure 3 illustrates a two-dimensional scatter visualization of all 12786 peaks identified from 941 signals, in which five classes are shown with different characters. The visualization is obtained by using t-Distributed Stochastic Neighbor embedding (t-SNE) [12,13]. This looks promising with regard to classification of four diseases and controls (WT). In Figure 4 we studied potential importance of 13 attributes applied to the entire peak data for the purpose of classification into five classes of the diseases and controls. In Table 1 means and standard deviations are given for 13 peak attributes class by class. The means gained mostly indicate clear differences between four diseases and controls which promises a good opportunity for their differentiation. Nevertheless, a few means of HCMM and HCMT are similar to each other.

2.2 Classification methods applied

Several classification methods were applied to identify test cases into five classes on the basis of peak data (the 13 aforementioned attributes): random forests [14-16], linear [17-19], quadratic [6,18,19] and Mahalanobis discriminant analysis [20], Naïve Bayes without kernel density estimation [21], with kernel

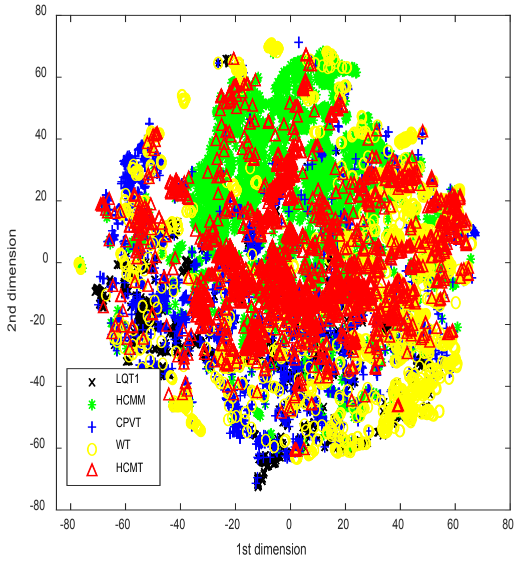

Figure 3: Two-dimensional visualization of calcium transient peak data of four diseases and controls (WT) computed with t-Distributed Stochastic Neighbour Embedding.

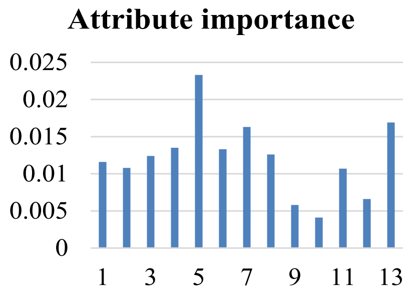

Figure 4: Importance analysis of 13 attributes. Each histogram value is the median of results generated by Matlab ReliefF algorithm applying nearest neighbor searching, when 12 runs were made with *k* (number of nearest neighbors) from {1, 3, 5, 7, 9, 11, 15, 21, 25, 31, 45, 61}. The higher value, the more important attribute.

Table 1: Means and standard deviations of peak attributes for diseases classes and controls (WT): left side amplitude A_l, right side amplitude A_r, left side duration D_l, right side duration D_r, maximum of the approximated first derivative *s*', absolute minimum of *s*', maximum of the second derivative *s*'' from the right peak side and its absolute minimum, peak area *R*, time difference Δ from peak maximum to maximum, duration d_l from the peak beginning to the location of the maximum of the first derivative of the left side, duration d_r from the peak maximum to the location of the absolute first derivative minimum of the right side and peak average width *w*.

Attributes	Diseases and controls (WT)				
	LQT1	HCMM	CPVT	WT	HCMT
A_l	170 ± 79	198 ± 92	229 ± 176	272 ± 170	199 ± 135
A_r	172 ± 80	200 ± 94	232 ± 176	275 ± 172	203 ± 138
D_l [s]	0.325 ± 0.178	0.272 ± 0.151	0.343 ± 0.199	0.492 ± 0.263	0.384 ± 0.187
D_r [s]	0.684 ± 0.404	0.476 ± 0.263	0.630 ± 0.433	1.039 ± 0.601	0.507 ± 0.332
max (*s*')	818 ± 472	1952 ± 888	1349 ± 1064	2131 ± 1276	1468 ± 1041
|min (*s*')|	509 ± 259	1030 ± 444	812 ± 541	927 ± 635	903 ± 479
max (*s*'')	1615 ± 1324	6002 ± 3189	2894 ± 2535	4465 ± 3386	4363 ± 3050
|min (*s*'')|	1208 ± 1432	3433 ± 3099	2106 ± 2709	3938 ± 4359	3405 ± 3645
R	57.7 ± 42.1	50.9 ± 43.5	84.7 ± 102.6	132.5 ± 114.8	61.8 ± 69.2
Δ [s]	1.17 ± 0.92	0.80 ± 0.50	1.13 ± 0.94	1.94 ± 1.58	1.02 ± 0.71
d_l [s]	0.216 ± 0.146	0.191 ± 0.122	0.212 ± 0.152	0.312 ± 0.198	0.290 ± 0.176
d_r [s]	0.154 ± 0.077	0.115 ± 0.072	0.144 ± 0.073	0.156 ± 0.145	0.107 ± 0.052
w [s]	0.400 ± 0.152	0.257 ± 0.114	0.382 ± 0.153	0.471 ± 0.259	0.280 ± 0.094

density estimation (normal kernel, Epanechnikov kernel, box kernel, and triangle kernel) [22], multinomial logistic regression [23,24], decision trees [25,26], K-nearest neighbor nearest searching (KNN) [25,27] with Chebychev metric, with cityblock (Manhattan) metric, with correlation measure, with cosine measure, with Euclidean metric, with Mahalanobis measure, with standardized Euclidean metric, with Spearman measure, and binary tree least square support vector machine (LS-SVM) [28-30] with the linear, quadratic, cubic and radial basis function (RBF) kernel. Note that all KNNs above were run with equal, inverse or squared inverse weighting. Moreover, the binary tree structure used with LS-SVM is described in Figure 5 where one class at a time is separated from the rest of the classes. The reason why WT is first separated in the tree structure is that we want to know whether or not the person has a cardiac disease overall before identifying the disease more detailed way. This chain of decisions simulates the actual situation what a physician encounters in reality. In the last node HCMM and HCMT are separated because they produce the same clinical phenotype.

When machine learning methods are used in practice, one cannot dismiss the parameter values since they have a huge impact on the final results. For the aforementioned classification methods, there are altogether four parameters to be examined more closely. Firstly, for KNN the k value is most important parameter. In this study, we tested the odd k values from 1 to 37. Only odd k values were tested to decrease the possibility of ties. Secondly, the LS-SVM classification method includes several tunable parameters. The number of parameters depends on the kernel selected. For all kernels the regularization parameter, C, is a common one. For the RBF kernel, the width of Gaussian (also known as σ) is a kernel specific parameter to be tuned. For both variables (C and σ) we used the same parameter value space {2-12, 2-11,…, 217} that led to a situation where we examined 30 values for the linear, quadratic and cubic kernels whereas with the RBF kernel we tested 900 (C, σ) combinations. Different parameter values were tested using grid-search and other, more advanced, methods for parameter value tuning such as the utilization of evolutionary computing are out of the scope of this paper. Thirdly, for random forests classifier the most crucial component is to select how many trees are included to a forest. In this study we examined the number of trees ranging from 1 to 100 with step size of 1. A forest can also consist of only one tree and due to the random aspect in random forests classifier, this tree differs from normal decision tree where all variables are used when constructing the tree. In random forests classifier only a subset of variables is used for constructing the tree in a forest.

Besides parameter values, an important issue to consider is the evaluation measures. There are no strict guidelines which evaluation measures should be used and they are always application and case specific choices. For this study we followed our earlier studies [4,5,6] and used sensitivity and accuracy as evaluation measures. Sensitivity can be computed for each class separately and it describes how well a specific class is recognized with respect to its class size in a test set. Accuracy instead explains the overall performance, i.e., what proportion of all signals is recognized correctly. Accuracy at a signal level is also used when finding the best parameter value for a classification method. The classification procedure described below is repeated with all parameter values tested and the highest accuracy determined the optimal parameter values.

Classification was performed by utilizing a variant of leave-one-out (LOO) method that is developed for the signal classification. The method is called leave-one-signal-out (LOSO). If we have N signals in a dataset, we have N rounds in classification just like in LOO. At each round in LOSO, the data from one signal in total is left for a test set and the rest of the data (data from N-1 signals) forms a training set. Here, we need to remember that

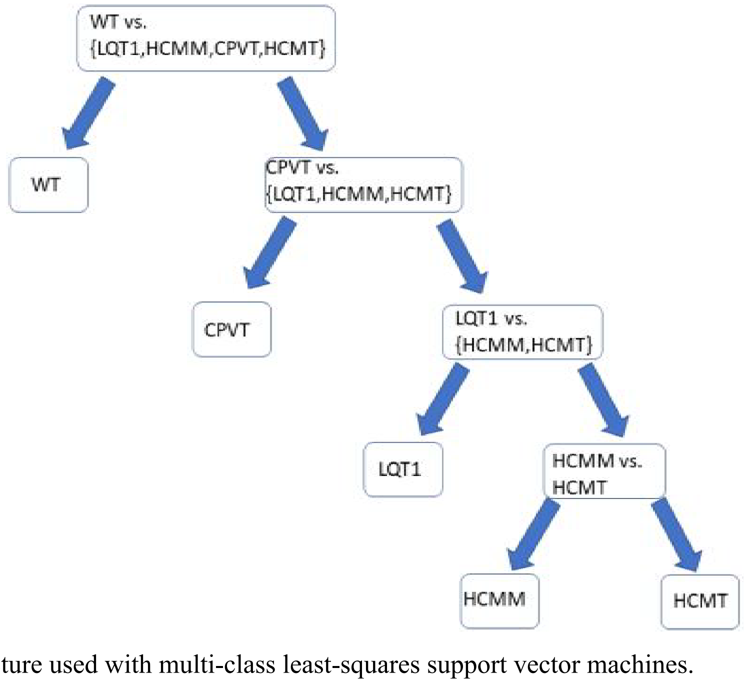

Figure 5: The binary tree structure used with multi-class least-squares support vector machines.

a data from one signal consists of several rows in observation matrix and each row represents the feature vector from one peak. Moreover, a classification method learns a model based on peak-level information and gives a prediction (class label) for each peak in a test set. In order to obtain a signal level classification result, we need to find the most frequent predicted class label within the peaks of a signal. For example, if we have a signal that has six peaks and belongs to class 1, the ground truth information for the peaks is (1,1,1,1,1,1) and the class label for signal is 1. Predicted class labels for the peaks within the signal can be, for instance, (1,1,2,1,2,1). Since there are four peaks with class label 1 and two peaks with class label 2, the predicted class label for the signal would be 1. This small example illustrates the signal level classification when the classification method gives predictions at peak-level.

When the most frequent predicted class label within a signal is determined, ties may occur. A tie can be, for instance, between classes 1, 2 and 3. Furthermore, a tie can be seen at peak-level as follows (1,3,2,3,1,2) when considering the earlier example. In order to solve a tie situation, we perform the following procedure. Find the classes included in a tie and find the corresponding number of peaks from a training set from these classes. Divide interval [0,1] with the same proportion to subintervals as there are peaks occurred in a training set with respect to the tied classes. Generate a random number from uniform distribution U(0,1) and find the subinterval where the random number belongs to. If we have a tie within three classes 1,2 and 3 like in earlier case and the corresponding number of peaks in a training set for classes 1,2 and 3 would be 20, 50 and 30, then the subintervals would be [0,0.2) for class 1, [0.2,0.7) for class 2 and [0.7,1] for class 3 respectively. If the random number generated would be 0.75, it belongs to sub-interval [0.7,1] and the predicted class label for the signal would be 3.

Seen from Table 2, true positive rates or sensitivities of the random forests that gave the best results from among all classifiers were 93% for LQT1, 87% for HCMM, 74% for CPVT, 74% for WT and 62% for HCMT. In Table 3 there are the detailed results of the random forests. From the row of HCMT in Table 3 we can see that HCMT is somewhat exposed to be predicted incorrectly to HCMM when 32 HCMT signals were incorrectly predicted to HCMM. On the other hand, 9 actually HCMM signals only were predicted incorrectly to HCMT disease. A probable cause is that HCMM is the majority class, much larger with 270 signals compared to that of HCMT with 149 signals. Another probable cause is that these two mutations of the same disease are not so dissimilar to each other than to other two diseases or controls. Both HCMM and HCMT are also mixed with controls (WT), when 18 actually HCMM and 20 actually HCMT signals were incorrectly classified into WT. Similarity and its opposite dissimilarity are on the basis of peak attributes computed. Cohen’s kappa for random forests was 0.73 (from interval [-1,1]) which is good being quite close to maximum 1.

Table 2: Sensitivities of four diseases and controls (WT) and classification accuracies in percent where *K* is the number of nearest neighbors that gave the best result reported, and *C* and *σ* are the control parameters for RBF kernel in binary tree LS-SVM. The best accuracy is given in Bold.

Classification method	**Sensitivity %**					**Accuracy %**
	LQT1	HCMM	CPVT	WT	HCMT	
Random forests, 21 trees	93.3	87.4	74.2	73.9	61.7	77.8
Decision trees	94.4	85.2	69.1	68.8	59.7	74.6
KNN, cityblock, equal, *K*=1	88.9	87	65.2	60.8	63.8	72.6
KNN, cityblock, inverse, *K*=3	88.9	85.6	64.4	65.8	64.4	73.1
KNN, cityblock, squared inverse, *K*=3	88.9	85.9	65.2	64.3	65.8	73.3
KNN, cosine, equal, *K*=1	78.9	83.3	66.1	64.8	59.7	71
KNN, cosine, inverse, *K*=1	78.9	83.3	66.1	64.8	59.7	71
KNN, cosine, squared inverse, *K*=3	81.1	83.3	66.5	65.3	58.4	71.2
KNN, Euclidean, equal, *K*=1	80	85.6	64.8	63.8	61.1	71.4
KNN, Euclidean, inverse, *K*=1	80	85.6	64.8	63.8	61.1	71.4
KNN, Euclidean, squared inverse, *K*=1	80	85.6	64.8	63.8	61.1	71.4
KNN, Mahalanobis, equal, *K*=1	83.3	84.1	63.1	62.8	60.4	70.6
KNN, Mahalanobis, inverse, *K*=3	82.2	87.8	62.2	62.8	61.7	71.5
KNN, Mahalanobis, squared inverse, *K*=7	83.3	88.9	64.4	59.3	66.4	72.5
KNN, standardized Euclidean, equal, *K*=1	80.3	85.6	64.8	63.8	61.1	71.4
KNN, standardized Euclidean, inverse, *K*=1	80.3	85.6	64.8	63.8	61.1	71.4
KNN, standardized Euclidean, squared inverse, *K*=1	80.3	85.6	64.8	63.8	61.1	71.4
Binary tree LS-SVM RBF kernel, *C*=32, σ=1	75.6	86.3	63.9	57.3	69.8	71

Table 3: Results of random forests in the form of a confusion matrix where the numbers in Bold of correctly classified signals (true positive) are located along the diagonal, the rows contain actual classes (four diseases and controls) and the columns those of classified (predicted). Numbers of incorrectly classified transient signals are those outside the diagonal. Percent values within the parentheses are counted along the rows.

	Predicted class by random forests for calcium transient signals				
True class	**LQT1 (%)**	**HCMM (%)**	**CPVT (%)**	**WT (%)**	**HCMT (%)**
LQT1	**84 (93)**	0 (0)	5 (6)	1 (1)	0 (0)
HCMM	0 (0)	**236 (87)**	7 (3)	18 (7)	9 (3)
CPVT	20 (9)	12 (5)	**173 (74)**	26 (11)	2 (1)
WT	1 (1)	18 (9)	23 (11)	**147 (74)**	10 (5)
HCMT	0 (0)	32 (22)	5 (3)	20 (13)	**92 (62)**

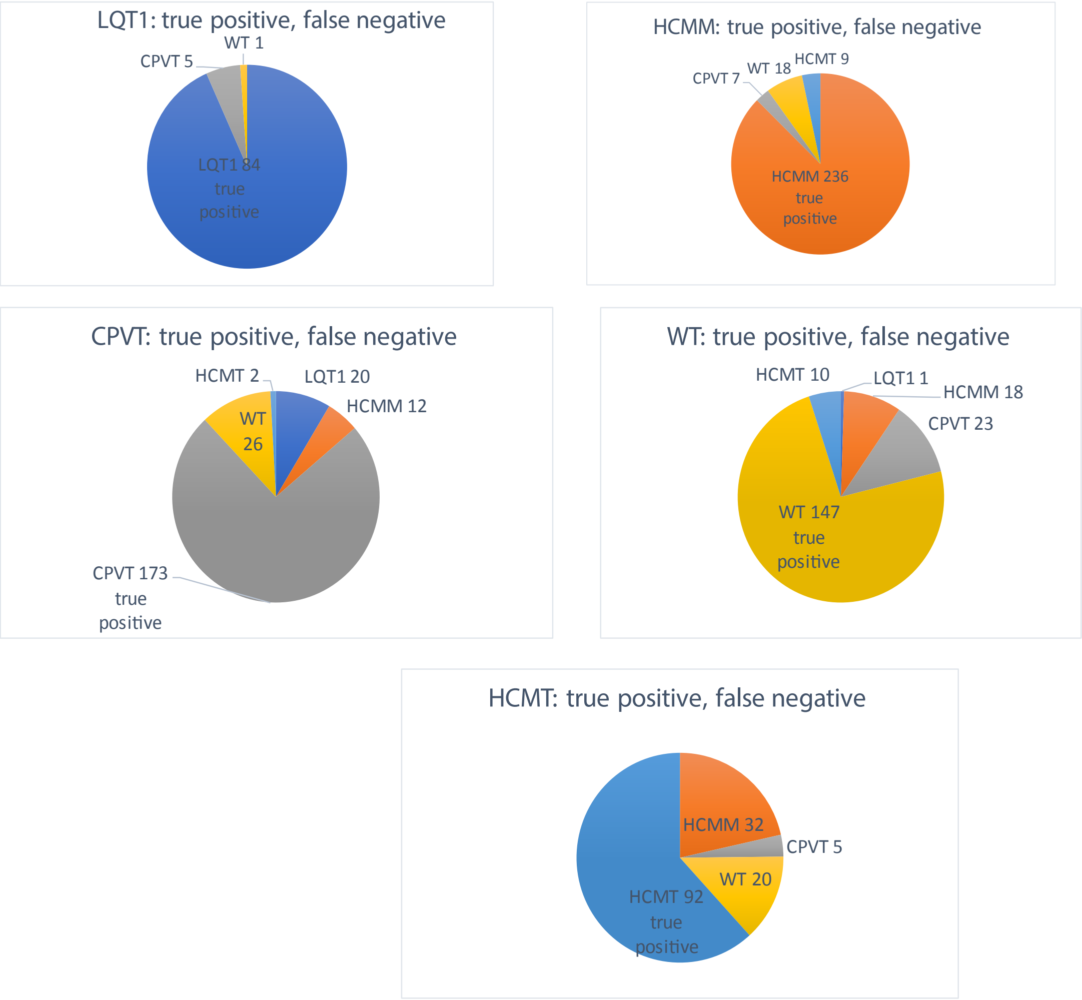

Figure 6: The five circles represent the four diseases and controls (WT, wild type). The circles show how many calcium transient signals were correctly classified (true positive) as the major sector in each of five classes and how many incorrectly classified (false negative) as smaller sectors to other classes. The best was LQT1 with 84 true positive and 6 false negative. The second best was HCMM with 236 true positive and 34 false negative. The next were CPVT with 173 true positive and 60 false negative and WT with 147 true positive and 52 false negative. The poorest was HCMT with 92 true positive and 57 false negative.

In order to clarify the contents of the confusion matrix in Table 3, Figure 6 shows pairwise class by class (diseases or controls) how classification failed on the basis of the numbers of the false negative. Figure 6 also shows true positive. i.e., it shows from which disease or controls those false negative signals were actually. Figure 7 shows the results from the other direction by presenting true positive and false positive. It shows to which classes (diseases or controls) some signals were incorrectly

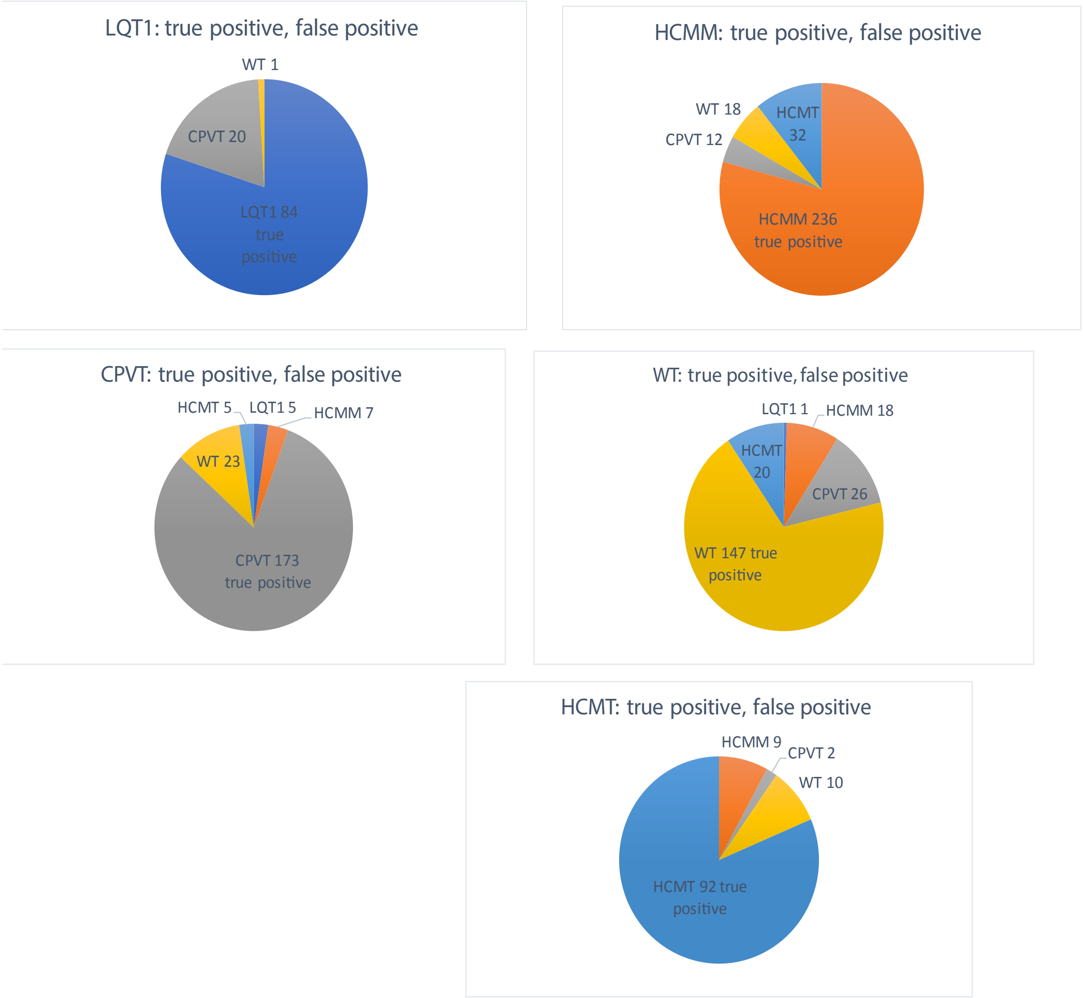

Figure 7: In each circle there is the major sector of true positive indicating the correctly classified class (disease or controls) being naturally the same as in the Fig. 6, whereas smaller sectors indicate incorrectly classified false positive with the class labels from which they actually originated.

classified. Thus, Figure 6 visualizes the results of Table 3 row by row, while Figure 7 does this column by column.

3. Results and Discussion

The sampling frequency of calcium transient signals were roughly 8 Hz for LQT1, 23 Hz for HCMM, 8 Hz for 55, 12 Hz for 93 and 23 Hz for 85 CPVT signals, 12 Hz for 40, 23 Hz for 93 and 33 Hz for 66 WT signals, and 14 Hz for 54 and 23 Hz for 95 HCMT signals. The increased sampling frequencies have been utilized along with the updated measuring system. Different sampling frequencies are, of course, not ideal. By considering the issue theoretically, the time difference ΔT of approximate sampling intervals 0.125 s and 0.030 s for the lowest and highest sampling frequencies f when f=1/T is approximately 0.095 s. While its variation is of type such as quantization noise, its average is ΔT/2 since its distribution is uniform, i.e., any value from the minimum to the maximum is equally probable. Now in theory this might be an average inaccuracy when the locations of a peak beginning, maximum or end were detected. Assuming that such inaccuracy might occur, for example, in the locations both beginnings and maxima of peaks for these two frequencies, total inaccuracy for the corresponding durations of peak left sides would be from minimum 0 s to maximum 2 (Δ T/2)=ΔT ≈ 0.095 s and approximately 0.048 on average. Looking at all six attributes directly dependent on time [s] in Table 1, we notice that time differences of the means of the five classes (four diseases and controls) are greater than 0.048 s for almost all pairs of classes except for attributes dl and dr. Nevertheless, subject to these extreme sampling frequencies there were 145 signals with 8 Hz and only 66 signals with 33 Hz, whereas there were 543 signals with 23 Hz being also the median sampling frequency used when representing 57.7% of all signals. In addition, 23 Hz was used for all other classes (270 HCMM, 85 CPVT, 93 WT and 95 HCMT signals) than LQT1. These two observations mean that any actual inaccuracy would be considerably less than the theoretical inaccuracy of 0.048 s computed above. Other seven attributes are only partially dependent on time.

Our results showed that with computational machine learning method HCM, LQT1 and CPVT diseases could be separated from each other by calcium transient signals with high accuracy.

In addition, HCM mutations HCMM and HCMT could also be separated from each other. This reinforces our previous findings and shows the possibility to discriminate genetic cardiac diseases and even different mutations by calcium transient profiles recorded from iPSC-CMs with machine learning classification methods.

Differentiation of five classes, i.e. four diseases and controls (WT), from each other was successful as the best classification accuracy of 77.8% generated by random forests for 941 signals in Tables 2 and 3 indicated. Previously, we obtained the best accuracy of 78.6% similarly by random forests for 527 calcium transient signals of LQT1, HCMM, CPVT and WT [6]. Now the numbers of HCMM and WT signals were greater and HCMT signals were also included. Although HCMT was somewhat slightly more difficult to differ from HCMM than it was with regard in all other pairs of classes, the entire results are still very high. The complexity of HCMT (149 signals) is natural, since HCMM (270 signals) was the majority class in the data. Possibly, it is even more influential that they are the mutations of the same disease.

We have previously shown that HCMT iPSC-CMs have more abnormal calcium transients than HCMM iPSC-CMs, however, abnormality types vary in both mutation types. These previous results together with other characterization methods suggested that abnormal calcium transients in HCM-CMs carrying different mutations may be caused by distinct mechanisms [31]. This study supports that observation by showing that machine learning method could be utilized to separate these two HCM mutations. This finding provides additional utilization of machine learning method for calcium transient signals to separate different disease mutations, which is important, since specific disease mechanisms of certain mutations may need mutation specific treatment.

4. Conclusion

The classification tests performed produced the interesting observation that it was still possible to get a very good classification accuracy, although the number of test signals was greatly increased and a set of new signals of HCMT were added, when compared to the results of our first article [6] related to the present research. These results reinforce our previous findings and encourage to continue and extend this research by increasing the number of HCM, LQT1 and CPVT patients as well as other inherited cardiac diseases, which lead to larger number of calcium signals to be analyzed. 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.

5. Acknowledgements

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

6. Conflict of Interest

There is no conflict of interests.

References

1. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell. 2007; 131: 861-872.
2. Moretti A, Bellin M, Welling A, Jung CB, Lam JT, Bott-Flügel L, et al. Patient-specific induced pluripotent stem-cell models for long-QT syndrome. N Engl J Med. 2010; 363: 1397-1409.
3. Penttinen K, Swan H, Vanninen S, Paavola J, Lahtinen AM, Kontula K, et al. Antiarrhythmic effects of Dantrolene in patients with catecholaminergic polymorphic ventricular tachycardia and replication of the responses using iPSC models. Plos One. 2015; 10: e0125366.
4. Juhola M, Penttinen K, Joutsijoki H, Varpa K, Saarikoski J, Rasku J, et al. Signal analysis and classification methods for calcium transient data of stem cell derived cardiomyocytes. Comput Biol Med. 2015; 61: 1-7.
5. Juhola M, Joutsijoki H, Penttinen K, Aalto-Setälä K. Machine learning to differentiate diseased cardiomyocytes from healthy control cells. Inf Med Unlocked. 2019; 14: 15-22.
6. Juhola M, Joutsijoki H, Penttinen K, Aalto-Setälä K. Detection of genetic cardiac diseases by Ca2+ transient profiles using machine learning methods. Sci Rep. 2018; 8: 9355.
7. Lee EK, Tran DD, Keung W, Chan P, Wong G, Chan CW, et al. Machine learning of human pluripotent stem cell-derived engineered cardiac tissue contractility for automated drug classification. Stem Cell Reports. 2017; 9: 1560-1572.
8. Heylman C, Datta R, Sobrino A, George S, Gratton E. Supervised machine learning for classification of the electrophysiological effects of chronotropic drugs on human induced pluripotent stem cell-derived cardiomyocytes. Plos One. 2015; 10: e0144572.
9. Maron BJ, Ommen SR, Semsarian C, Spirito P, Olivotto I, Maron MS. Hypertrophic cardiomyopathy: present and future, with translation into contemporary cardiovascular medicine. J Am Coll Cardiol. 2014; 64: 89-99.
10. Mummery C, Ward-van Oostwaard D, Doevendans P, Spijker R, Van den Brink S, Hassink R, et al. Differentiation of human embryonic stem cells to cardiomyocytes: role of coculture with visceral endoderm-like cells. Circulation. 2003; 107**:** 2733-2740.
11. Kujala K, Paavola J, Lehti A, Larsson K, Pekkarinen-Mattila M, Viitasalo M, et al. Cell model of catecholaminercig polymorphic ventricular tachycardia reveals early and delayed after depolarizations. Plos One. 2012; 7: e44660.
12. Van der Maaten LJP. Accelerating t-SNE using tree-based algorithm. J Mach Learn Res. 2014; 15: 3221-3245.
13. Van der Maaten LJP, Hinton GE. Visualizing high-dimensional data using t-SNE. J Mach Learn Res. 2008; 9: 2579-2605.

14. Breiman L. Random forests. Mach Learn. 2001; 45: 5-32.

15. Cutler DR, Edwards Jr TC, Beard KH, Cutler A, Hess KT, Gibson J, et al. Random forests for classification in ecology. Ecology. 2007; 88: 2783-2792.

16. Genuer R, Poggi J-M, Tuleau-Malot C. Variable selection using random forests. Pattern Recogn Lett. 2010; 31: 2225-2236.

17. Balakrishnama S, Ganapathiraju A. Linear discriminant analysis – A brief tutorial. Technical report, Institute for Signal and Information Processing 1998; 1-8.

18. Cios KJ, Pedrycz W, Swiniarski RW, Kurgan LA. Data Mining: A Knowledge Discovery Approach. Springer, New York, 2007.

19. Izenman AJ. Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning. Springer, New York. 2008.

20. Bohling G. Classical normal-based discriminant analysis. Technical report, Kansas Geological Survey. 2006: 1-24.

21. Aly M. Survey on multiclass classification methods. Technical Report, Cal Tech. 2005: 1-9.

22. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning – Data Mining, Inference, and Prediction, 2nd ed, Springer. 2009.

23. Agresti A. Categorical Data Analysis, John Wiley & Sons. 1990.

24. Kwak C, Clayton-Matthews A. Multinomial logistic regression. Nurs Res. 2002; 51: 404-410.

25. Duda RO, Hart PE, Stork DG. Pattern Classification, 2nd ed, John Wiley & Sons. 2001.

26. Loh W-Y. Classification and regression trees. Wiley Interdiscip. Review: Data Min. Knowl. Discov. 2001; 1: 14-23.

27. Cunningham P, Delany SJ. k-nearest neighbor classifiers. Technical Report UCD-CSI-2007-4. 2007; 1-17.

28. Joutsijoki H, Haponen M, Rasku J, Aalto-Setälä K, Juhola M. Machine learning approach to automated quality identification of human induced pluripotent stem cell colony images. Comput Math Meth Med. 2016; 3091039: 15.

29. Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett. 1999; 9: 293-300.

30. Suykens JAK, Vandewalle J. Multiclass least squares support vector machines. In: Proceedings of the International Joint Conference on Neural Networks. 1999; 900-903.

31. Ojala M, Prajapati C, Pölönen RP, Rajala K, Pekkanen-Mattila M, Rasku J, et al. Mutation-specific phenotypes in hiPSC-derived cardiomyocytes carrying either myosin-binding protein C or α-tropomyosin mutation for hypertrophic cardiomyopathy. Stem Cells Int. 2016; 1684792.