Informatics in Medicine Unlocked 14 (2019) 15–22

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

Machine learning to differentiate diseased cardiomyocytes from healthy control cells

Martti Juhola[a,*], Henry Joutsijoki[a], Kirsi Penttinen[b], Katriina Aalto-Setälä[b,c]

[a] Faculty of Information Technology and Communication Sciences, Tampere University, Finland
[b] Faculty of Medicine and Health Technology, Tampere University, Finland
[c] Heart Center, Tampere University Hospital, 33520, Tampere, Finland

ARTICLE INFO

Keywords:
Calcium transient profiles
Genetic cardiac diseases
Machine learning
Differentiation of the diseased from controls

ABSTRACT

Human induced pluripotent stem cell-derived cardiomyocytes (iPSC-CMs) have been shown to be useful to improve techniques that are developed for the study of cardiac disease. Abnormalities in Ca^{2+} transients are commonly present in iPSC-CMs derived from individuals with a cardiac disease. We previously observed that Ca^{2+} transient signals of healthy CMs can be distinguished from transients of CMs derived from individuals having different genetic cardiac diseases. Machine learning was used to distinguish different diseases from each other as well as from controls. We wanted further to investigate whether we are able to separate iPSC-CM Ca^{2+} signals of any genetic cardiac disease as one group from those of healthy individuals by utilizing machine learning methods. A total number of 593 CM transient signals from healthy individuals and from patients were analyzed. We obtained a best classification accuracy of 87% between the disease group and controls. This finding provides evidence that machine learning methods are efficient for identifying iPSC-CMs derived from individuals with a disease phenotype, and that iPSC-CMs may be useful to identify individuals at risk for a cardiac event.

1. Introduction

Genetic cardiac diseases present a wide range of symptoms, ranging from completely asymptomatic to severe arrhythmias, and even sudden cardiac death [1,2]. Additionally, most if not all of these diseases have an increased risk for arrhythmia, in addition to structural or other cardiac abnormality, e.g., in various cardiomyopathies. If the mutation causing disease is known in the family, it is easy to focus on the mutation carriers for follow-up and primary prevention of potential arrhythmias. However, this is often problematic when the disease is presented by sudden death in the family, but no mutations are found. In such situations, induced pluripotent stem cell (iPSC) -derived cardiomyocytes may provide a useful alternative to predict arrhythmic risk, and to identify those family members with increased risk of clinical symptoms including arrhythmias.

On a cellular level, cardiac functionality can be studied with the help of CMs differentiated from human pluripotent stem cells [3,4]. The induced pluripotent stem cell (iPSC) technology offers a way to reprogram differentiated cells back to the pluripotent state – and, therefore, it is a useful tool for studying the pathophysiology of various disorders and drug responses in human cells [5]. Additionally, cellular differentiation and maturation can be studied with these cells [6].

Thus far, iPSC-CMs have successfully been used to model genetic cardiac diseases such as catecholaminergic polymorphic ventricular tachycardia (CPVT) [3,4,7–12], long QT syndrome (LQT) [13–16] and hypertrophic cardiomyopathy (HCM) [17–19]. iPSC-derived CMs have revealed considerable abnormalities and diversity in intracellular Ca^{2+} cycling features compared to healthy control CMs. Ca^{2+} cycling plays an important role in cardiac functionality by linking electrical activation and contraction, and the characterization of Ca^{2+} cycling is vital in order to facilitate investigations of cardiac disorders and dysfunctions, as well as to study disease management with different compounds.

Heretofore, machine learning has rarely been applied to the data associated with induced pluripotent stem cell-derived cardiomyocytes. Machine learning has however been applied to the mechanistic action of cardioactive drugs [20]. We demonstrated in previous articles that the separation of calcium transient signals of abnormally and normally grown cardiomyocytes can be accurately done with machine learning [21,22]. In these papers, abnormality is defined as deformed Ca^{2+} peak forms varying in amplitudes (sizes) and durations of Ca^{2+} transients and normality as harmonious transients of approximately the same size and form throughout entire calcium transient signals. To the best of our knowledge, our recent study [23] was the first one in which different genetic cardiac diseases were separated according to their calcium

* Corresponding author.
E-mail address: Martti.Juhola@tuni.fi (M. Juhola).

https://doi.org/10.1016/j.imu.2019.01.006
Received 17 December 2018; Received in revised form 16 January 2019; Accepted 31 January 2019
Available online 02 February 2019

transient signals by using classification performed with machine learning methods.

In the present research, visually normal and abnormal Ca^{2+} transient signals and peak variables of three genetic cardiac diseases and healthy control CMs were used. Disease-specific CMs were generated from patients suffering from CPVT, an exercise-induced malignant arrhythmogenic disorder [3,8], LQT type 1, an electric disorder of the heart that predisposes patients to arrhythmias, and HCM, a disorder that affects the structure of heart muscle tissue leading to arrhythmias and progressive heart failure [19]. All of the transients of diseased cells were pooled and compared to all of the transients obtained from control cells, and different machine learning algorithms were designed and used to analyze this technology to automatically distinguish the groups.

2. Materials

The current study was approved by the Ethics Committee of Pirkanmaa Hospital District in establishing, culturing and differentiating human iPSC lines (R08070). The patient-specific iPSC lines were established and characterized as described earlier, as well as the CM differentiation and dissociation of beating areas [22]. The studied cell lines included six CPVT lines generated from CPVT patients carrying cardiac ryanodine receptor (RyR2) mutations, four HCM cell lines generated from HCM patients carrying either α-tropomyosin (TPM1) or myosin-binding protein C (MYBPC3) mutations, two LQT type 1 cell lines generated from patients carrying potassium voltage-gated channel subfamily Q member 1 (KCNQ1) mutations, and one cell line generated from a healthy control individual. Thus, there were 13 subjects altogether.

Ca^{2+} imaging was conducted in spontaneously beating, 4 μM Fura-2 AM (Invitrogen, Molecular Probes) or 4 μM Fluo-4 AM (Life Technologies Ltd) loaded dissociated CMs as described earlier [3]. During the measurements, CMs were perfused with 37 °C HEPES based perfusate consisting of (in mM) 137 NaCl, 5 KCl, 0.44 KH2PO4, 20 HEPES, 4.2 NaHCO3, 5 D-glucose, 2 CaCl2, 1.2 MgCl2, and 1 Na-pyruvate (the pH was adjusted to 7.4 with NaOH). 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 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 Ca^{2+} analysis, regions of interest were selected for spontaneously beating cells and background noise was subtracted before further processing. Each Ca^{2+} signal corresponded to a recording from one cell.

3. Data computed from Ca^{2+} transient signals

Human induced pluripotent stem cell-derived CMs were the data source from which cycling Ca^{2+} transient signals were obtained. Data used in the computation were based on the peaks of Ca^{2+} transient signals. Cycling peaks were recognized, and data variables or features were extracted from every peak. Previously Ca^{2+} transient signals were categorized using our recognition algorithm [22], which classified them into either normal type or abnormal signal type on the basis of normal or abnormal peaks of the signals, and where we observed that it was possible to separate normal from abnormal signals up to the accuracy of approximately 90% when compared to a human expert's classification decisions.

Fig. 1 presents 10 s segments of four signals as examples. The signals were short, their durations being from 7.7 s to 46.5 s and 19.0 s on average. An entire signal was determined to be abnormal if even a single peak was observed as abnormal. In the current research, we did not differ normal and abnormal signals from each other, but advanced to separate Ca^{2+} transient signals of diseased induced pluripotent stem cell-derived CMs from those of control subjects. It was interesting to study whether disease transient signals can be differentiated from those of controls, although both classes contained both normal and abnormal signals. The number of the abnormal control signals was only 12.6% of all control signals. The disease transient signals originated from the group of the three above-mentioned diseases: LQT1, HCM and CPVT. These were used jointly as the disease class. The other class was formed by the signals of the controls (wild type, WT).

The data used comprised 394 disease transient signals and 199 control transient signals. These contained, respectively, 179 normal and 215 abnormal signals, and 174 normal and 25 abnormal signals. It is noticeable that there were only relatively few abnormal signals in controls, since these were far more infrequent compared to those of disease CMs.

Sampling of the transient signals contained three different frequencies, because the data were recorded at different times and the sampling frequency was increased in the meantime. The approximate sampling frequencies were 8 Hz, 11 Hz and 23 Hz. In this order, 35%, 26% and 39% of the diseased signals were recorded, and, correspondingly, 5%, 15% and 80% of the control signals.

In order to detect individual peaks of a transient signal, values of its first derivative were computed in short sequential segments, where slopes of linear regression computed with sequential signal segments of a few samples were used to estimate first derivative values (Fig. 2). To determine the beginning of a peak, its first derivative values had to remain smaller than a small positive threshold value determined experimentally during a few sequential slope values. Thereafter, slope values became greater while proceeding forward along a typically steep left side of a peak producing large positive first derivative slope values. Next a peak maximum or top was found when slope values again decreased less than the positive threshold value. After the maximum, first derivative slope values changed negative along the decreasing right side of a peak (Fig. 2). Ultimately, the end of a peak was encountered when the first derivative slope values again increased close to zero. The detailed procedure for the peak detection was introduced in our previous research [22]. However, oscillations of very small amplitudes were not accepted as valid peaks in a signal, as follows. After the removal of a possible linear trend in a signal, the amplitude of large peaks in a signal was estimated as a difference from the average of the highest sample (amplitude) values (15% of all) to the lowest sample values in the current signal. Such peak candidates that had the amplitude of the left or right side of a peak less than approximately 8% from the above amplitude estimate of the large peaks were not accepted as peaks, but were suspected as being probable noise [22]. The numbers of the peaks extracted from the signals varied from 1 to 61, and were only 12.8 on average.

After the recognition of the peaks from the transient signals, values of 12 variables for every peak were computed as follows. First, the amplitudes of the left and right side of a peak were computed - see Fig. 2. Second, the durations of both peak sides were computed from locations *a* to *c* and from *c* to *e*. Third, the maximum of the first derivative from the left side of a peak and the absolute minimum of the first derivative from the right side were computed. Fourth, the maximum and absolute minimum of the second derivative were computed from the right side only. The left side was not now applied, because frequently these were too short (as to the number of samples) for second derivative values to have been calculated. Fifth, the surface determined by a peak curve and a line from the beginning to the end of a peak was computed. Sixth, the duration (time difference) from the maximum at location *c* to the maximum of the preceding peak was computed, or if the current peak was the first peak of a signal, the duration was calculated from the signal beginning. Seventh, the duration (time difference) from peak beginning *a* to location *b* of the first derivative

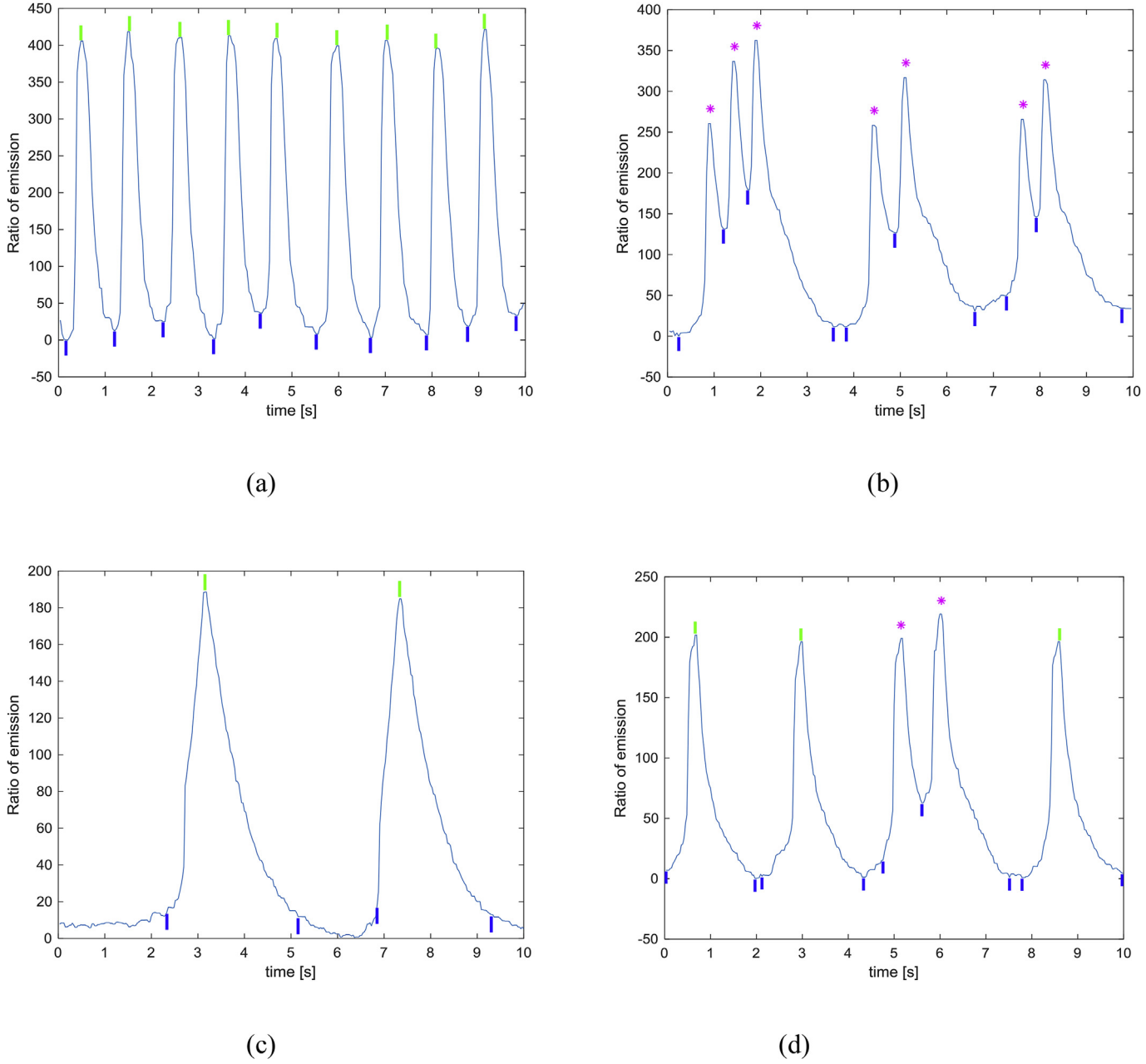

Fig. 1. (a) A 10 s segment of a normal CPVT signal in which the peaks recognized to be normal are marked with green bars, and also the beginning and end of every peak marked. (b) An abnormal CPVT signal in which all peaks were recognized as deformed to be abnormal and marked with stars. (c) A 10 s segment of a normal control transient signal. (d) An abnormal control transient signal, where two abnormal peaks were marked with stars. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

maximum of the left peak side and the duration from location *c* of the peak maximum to location *d* of the first derivative minimum of the right peak side were computed.

To visualize the data, in other words, variable values computed from 5290 recognized peaks of the disease transient signals and 2291 peaks of the control transient signals, after the normalization of the data Stochastic Neighbor Embedding algorithm with the Euclidean distance measure in MATLAB was used to present the data in two dimensions. This is depicted in Fig. 3. When a great part of the cases in two different classes are apart from each other, it is possible that this predicts a successful classification for the classes.

The means and standard deviations of all 12 variables are presented in Table 1. Considerable differences for the means of two classes for every variable are seen, which may denote a favorable classification chance between the disease and control transient signals.

Next we evaluated how efficient the 12 variables are to separate or classify the two classes. We ran the reliefF algorithm in MATLAB for our data. It functions on the basis of applying a nearest neighbor searching method, in order to measure the differentiation power or weight, variable by variable. We chose nine *k* values for the numbers of nearest neighbors, control parameter of the algorithm. They were 3, 5, 7, 9, 11, 15, 21, 25 and 31, where odd values only were used to prevent possible ties (equal numbers from the two opposite classes) during nearest neighbor searching. For each peak variable, the median of weights given by the reliefF algorithm for results of nine runs (Fig. 4) was computed. The positive weights mean that all variables are able to separate the two classes. Weights are relative and they express ranking of variables for differentiation of classes. For all nine runs, variables 3, 4, 5

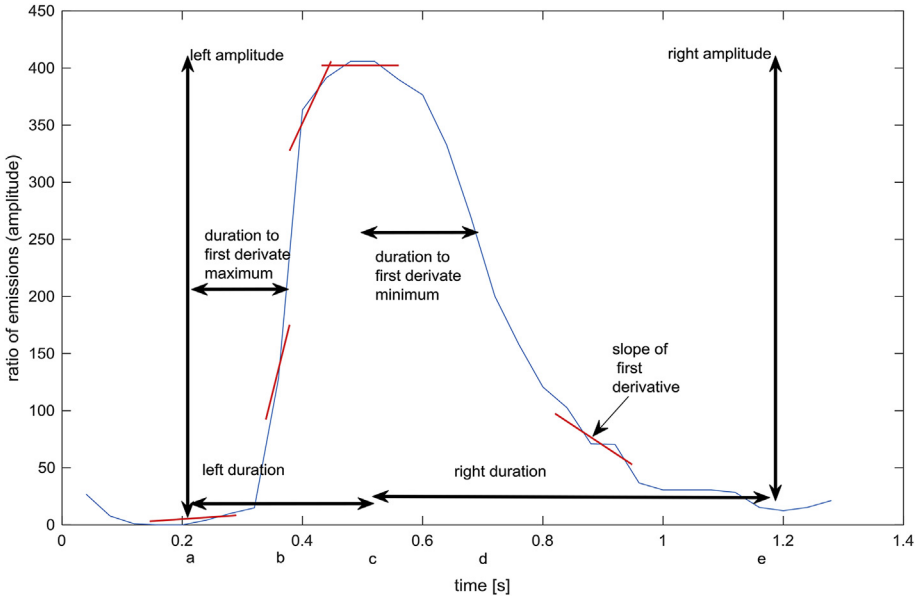

Fig. 2. The peak is the first one from the signal in Fig. 1(a). The short red lines visualize imaginary slopes (upward is positive, downward negative and horizontal zero slope value) of the first derivative curve of the signal during the current peak. The peak begins from location *a*; its maximum is at *c* and end at *e*. Location *b* is for the maximum of the first derivative and *d* for its minimum. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

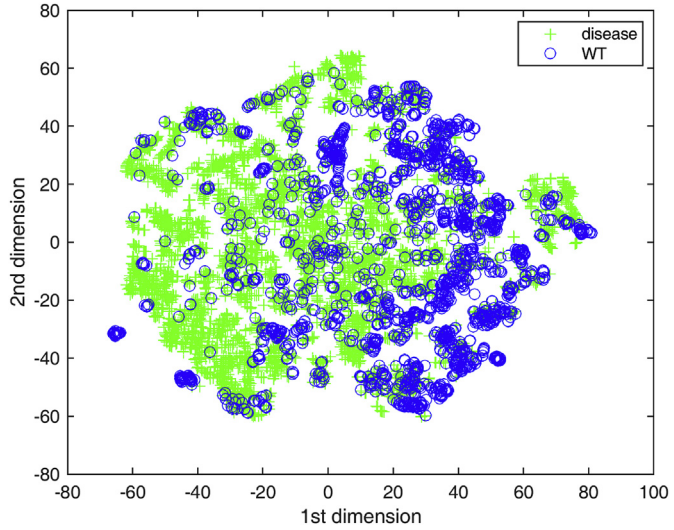

Fig. 3. The visualization of the peak variable values of disease and control (WT) transient signals in two dimensions.

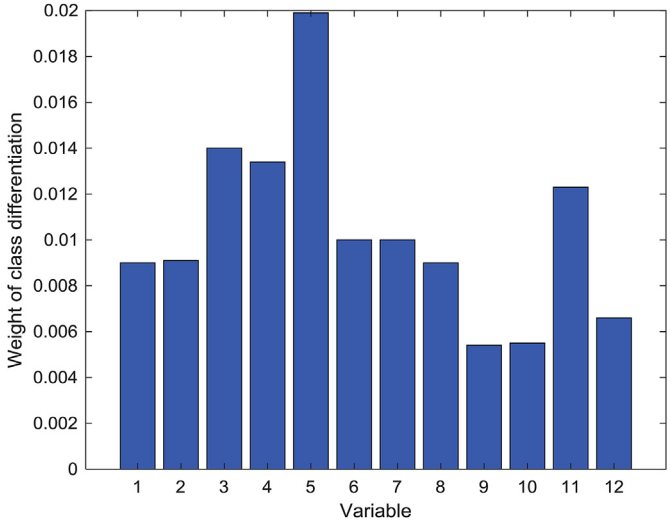

Fig. 4. Medians of weights were calculated with the reliefF algorithm for the 12 variables of the current data by using 9 different *k* values (number of nearest neighbors). Variables 3, 4, 5 and 11 are the most efficient and variables 9, 10 and 12 the least efficient for the differentiation of the disease transient signals from those of controls.

and 11 obtained the greatest or best weights and variables 9, 10 and 12 resulted in the smallest or poorest instances to separate the two classes. Instead, variables 1, 2, 6, 7 and 8 were between the best and poorest, and their ranking varied slightly among the nine test runs, when their

Table 1
Means and standard deviations of 12 variables for 5290 peaks of the disease transient signals and 2291 peaks of the control transient signals.

Variable number	Peak variable	Disease transient signal peaks	Control transient signal peaks
1	Peak left side amplitude	201.3 ± 134.4	272.5 ± 170.2
2	Peak right side amplitude	203.4 ± 135.4	275.4 ± 171.8
3	Left duration [s]	0.31 ± 0.181	0.492 ± 0.263
4	Right duration [s]	0.597 ± 0.394	1.039 ± 0.601
5	Maximum of left side first derivative	1348 ± 985	2131 ± 1276
6	Absolute minimum of right side first derivative	780 ± 496.33	927 ± 635
7	Maximum of right side second derivative	3397 ± 3110	4465 ± 3386
8	Absolute minimum of right side second derivative	2116 ± 2561	3938 ± 4359
9	Peak area	68 ± 76	132 ± 115
10	Duration from peak maximum to preceding one (or beginning) [s]	1.039 ± 0.858	1.944 ± 1.58
11	Duration from peak beginning [s] to left side maximum of first derivative	0.201 ± 0.14	0.312 ± 0.198
12	Duration from peak maximum [s] to right side minimum of first derivative	0.138 ± 0.075	0.156 ± 0.145

weights were approximately equal as seen in Fig. 4. In any case, on the basis of these results as well those in Table 1, all 12 variables were found to be useful for the differentiation of disease and control calcium transient signals from each other.

4. Classification methods used and design of experiments

We used a wide collection of classification algorithms in our study, ranging from traditional methods to state-of-the-art methods. Since our dataset consists of a true class label defined by the human expert for each signal, we concentrated only on supervised classification methods and, hence, semi-supervised and unsupervised methods were out of the scope of this paper. Methods used in our study were mostly the same as used in our previous study [23]. However, in Ref. [23] the classification task was multi-class by nature, whereas in this paper the classification task is a two-class problem. We do not present the detailed description about the actual classification methods, but a reader can find thorough descriptions about the algorithms from the given references. Compared to Ref. [23], there were now two novel peak variables: variables 11 and 12 (Table 1).

As a first classification algorithm, we applied the *k* Nearest Neighbor (*k*NN) searching method [24–26] that is one of the earliest classification algorithms and most used. The performance of the *k*NN algorithm depends mainly on three factors: *k* value, distance measure, and distance weighting scheme. These factors are data-dependent and for each dataset, a suitable combination must be searched separately. For our study, we selected *k* values of 1, 5, 7, 11, 13 and 17 to be examined, and here we followed the principle used in Ref. [23]. We selected eight distance measures to be tested: Chebyshev, cityblock (Manhattan), correlation, cosine, Euclidean, Mahalanobis, standardized Euclidean and Spearman. We performed classification with three distance weighting schemes: no weighting (weighting equal to 1), reciprocal, and squared reciprocal with respect to distance.

The second wholeness used was discriminant analysis based algorithms. Discriminant analysis covers various variations, and from them we applied linear discriminant analysis [27,28], quadratic discriminant analysis [28,29], and Mahalanobis distance based discriminant analysis [30]. When moving to probability based algorithms, naïve Bayes classifier [24,31,32] cannot be dismissed. Naïve Bayes is a classical widely used method in many applications. It can be used with or without kernel density estimation [24,31]. In this paper we used the naïve Bayes algorithm in both senses. When kernel density estimation was used, we examined Gaussian, box, triangle and Epanechnikov kernels. We also applied the naïve Bayes classifier without kernel density estimation when the normal distribution assumption over the dataset is expected. Besides naïve Bayes classifier, *k*NN and discriminant analysis based classification methods, we used multinomial logistic regression [33,34] which returns to logistic regression [31,32] in two-class tasks.

Decision tree-based solutions are commonly used alternatives in machine learning tasks. Their advantages are easy interpretation and computational efficiency. These issues are important to take into account when considering the end-users of our application, who are persons not experts in the machine learning area. In our study, we investigated CART [25,35] algorithm and Random Forests [36–38]. For Random Forests we varied the number of trees from 1 to 100.

The Support Vector Machine [39] has gained great popularity since the early 1990s, and has been used in various applications. However, in our study, we decided to use a variant of SVM called the least squares support vector machine (LSSVM) [40–42] which differs from traditional SVM in such a way that LSSVM solves a system of linear equations instead of a quadratic optimization problem. The performance of LSSVM is heavily dependent on the selection of a kernel function and (hyper)parameter values. Hence, it is always necessary to perform a thorough search of (hyper)parameter values in order to ensure the best possible result. A common parameter for all kernel functions in LSSVM is C, also called box constraint. We selected the linear, quadratic, 3rd degree of polynomial kernel and the RBF kernel to be examined in our paper. The parameter value space for the box constraint and hyperparameter σ (the width of Gaussian function) is $\{2^{-12},2^{-11},\ldots,2^{16},2^{17}\}$. By this means polynomial kernels were tested on 30 parameter values and the RBF kernel on 900 (C,σ) combinations.

The classification was performed based on the leave-one-signal-out (LOSO) procedure, which is a modification compared to the leave-one-out method. In signal classification, a noticeable detail must be remembered. Variables are determined from peaks, and a signal consists of one or more peaks. Thus, the data gained from one signal usually includes several rows in an observation matrix. When defining training and test sets, one needs to ensure that the whole data from the signal is in either the training or test set. Signal data must not be split into two, such that one part is in training set and the other is in test set. In LOSO, the peak-based data from each signal in turn forms a test set, and the rest of the data are in the training set.

When we train a computational model based on a classification method, we must remember that the training phase of an algorithm is performed on peak-based data, not on signal level data. Hence, a classifier learns its model based on peaks and not on signals. After training of a classifier, we give the test set as an input to the classifier. Then the classifier gives a predicted class label for each peak in a test set. In this stage, results are at peak level. However, since the aim of this paper is signal classification, we need to transform the peak level results into a signal level result. This is done by taking a mode from the predicted class labels for peaks in a test set. We can do this because we use the LOSO procedure, where a test set covers data only from one signal. However, mode is not always unambiguously defined and a tie may occur. In our paper we had only two classes, so a tie could occur with only two classes. If a tie occurred, we solved the problem in the following way.

1. Extract the training data of classes C_1 and C_2 occurring in a tie, from the training set.
2. Find the proportions $P_1=(|C_1|/(|C_1|+|C_2|))100\%$ and $P_2=(|C_2|/(|C_1|+|C_2|))100\%$ where $|.|$ is the size of a set. Hence, an interval $[0,P_1]$ is for the class C_1 and interval $(P_1,100]$ is for the class C_2.
3. Generate a random number R from the uniform distribution U(0,1).
4. If R*100% belongs to interval $[0,P_1]$, select C_1 as final class label for the signal. Otherwise, select C_2.

After finding a predicted class label for each signal in a dataset, we can compare the predicted class label with the true label and construct a confusion matrix. From the confusion matrix, we can evaluate different kinds of measures which describe how well the classification has succeeded. For our study, we selected accuracy ((TP + TN)/(TP + TN + FN + FP)), true positive rate (TP/(TP + FN)) for diseases and true negative rate (TN/(TN + FP)) for controls. For the classification methods which require parameter tuning, we repeated LOSO with all parameter values examined, and selected a parameter value (combination) that achieved the highest accuracy.

5. Classification results of disease or control transient signals

The main target of the research was to study how efficiently two transient signal groups can be differentiated from each other. For this purpose, several classifiers were implemented as described above. Their results are presented in the following.

Classification results are shown in Tables 2–4, in which true positive rates (sensitivity) correspond to disease transient signals, and true negative rates (specificity) to control signals. Accuracy equals the sum of true positive and negative cases divided by the number of all cases. Now *k* nearest neighbor searching with cityblock (Manhattan) metric yielded the best accuracy results of 86.0% in Table 2. In Table 3, *k* nearest neighbor searching with Euclidean metric and squared inverse weighting was the best method, obtaining a 84.5% level. In Table 4, the

Table 2
Classification results of k nearest neighbor (kNN) searching, with different metrics or measures with the best k value.

Classification method	True positive rates of diseases %	True positive rates of controls %	Accuracy %
kNN with Chebychev metric and equal weighting, $k = 1$	87.3	69.3	81.3
kNN with Chebychev metric and inverse weighting, $k = 1$	87.3	69.3	81.3
kNN with Chebychev metric and squared inverse weighting, $k = 5$	86.5	71.4	81.5
kNN with cityblock metric and equal weighting, $k = 1$	91.1	75.9	**86.0**
kNN with cityblock metric and inverse weighting, $k = 1$	91.1	75.9	**86.0**
kNN with cityblock metric and squared inverse weighting, $k = 1$	91.1	75.9	**86.0**
kNN with correlation measure and equal weighting, $k = 1$	89.3	67.8	82.1
kNN with correlation measure and inverse weighting, $k = 5$	89.3	71.4	83.3
kNN with correlation measure and squared inverse weighting, $k = 5$	89.3	71.9	83.3
kNN with cosine measure and equal weighting, $k = 1$	86.8	71.4	81.6
kNN with cosine measure and inverse weighting, $k = 5$	87.8	74.4	83.3
kNN with cosine measure and squared inverse weighting, $k = 7$	89.3	72.9	83.8

support vector machine (LSSVM) with radial basis function (RBF) kernel having an accuracy of 84.7% and random forest with 87.4% were the best techniques. Overall, the classification of data into two classes was very successful.

6. Discussion

This study was aimed at investigating whether machine learning could, in general, separate healthy cardiomyocytes from diseased ones. The phenotype of the iPSC-derived CMs was determined by Ca^{2+} imaging. Both control and diseased cardiomyocytes contained cells with normal beating, as well as those with abnormal beating behavior. With machine learning, very high classification accuracy values (up to 87.4%) were obtained to distinguish control and diseased cells, despite both having mixed CM populations (containing both normal and abnormal CMs), i.e., to distinguish CMs derived from patients carrying a mutation for a cardiac disease from control CMs.

The iPSC technology has revolutionized the study of genetic cardiac diseases [43]. It enables the investigation of patient- and mutation-specific cells in order to understand the disease pathophysiology of interest, as well as to provide a platform to study drug responsiveness in a personalized way [4]. However, CMs derived from iPSCs obtained with current differentiation protocols still have several problems. They are first of all immature cardiomyocytes [44]. In addition, they present all types of cardiomyocytes including atrial, ventricular and pacemaker cells. These issues make it problematic to produce and determine a disease phenotype in a reproducible way. In our current study, these limitations were evident. The CMs were immature, and both our control CMs as well as the diseased CMs were mixed cell populations and with Ca^{2+} transient signals, e.g., ventricular or atrial cells cannot be distinguished from each other. However, despite these problems, our machine learning procedure was successful to differentiate control CMs from diseased CMs, suggesting the presence of characteristics of healthy or diseased cells already in the fatal state and common for all types of CMs.

Due to above mentioned problems with iPSC-derived CM, control cells also included abnormal Ca^{2+} transients. In our study, 12.6% of control cells presented various types of abnormalities. However, the amount of abnormal transient signals was much greater (54.6%) in CMs carrying any mutation for various genetic cardiac diseases including genetic arrhythmias and cardiomyopathies. We demonstrated earlier [23] that it is possible to differentiate genetic cardiac diseases from each other based on machine learning techniques. In the current study, the aim was to collect more transient signals of control CMs, to decrease the difference between the smaller signal number of control CMS and the greater signal number of diseased CMs compared to the earlier situation [23], and pool all diseased ones as one group, to analyze whether this could be separated from control cells containing also both abnormally and normally beating cells. Using different algorithms and methods for machine learning, we were able to produce paradigms with high a classification accuracy of up to 87.4%, suggesting that this procedure could have potential use in clinical applications in the future.

7. Conclusions

Genetic cardiac diseases are clinically often problematic. First, the disease phenotype is variable even within a single family [45]. Additionally, it is still common despite advances in molecular genetics that the mutation causing the disease in the family is not known. In this situation, it is impossible to distinguish those who are potentially at risk of developing the disease phenotype, which family members should be regularly checked clinically, and who should be advised for lifestyle restrictions or preventive medication. The results obtained in this study are potentially promising to identify individuals at risk. iPSC-derived CMs carrying a disease causing mutation can accurately be separated from cells derived from healthy individuals, and thus potentially iPSC-derived cardiomyocytes combined with machine learning algorithms could in the future be used also clinically to identify individuals at risk,

Table 3
More classification results of k nearest neighbor (kNN) searching, with different metrics or measures with the best k value.

Classification method	True positive rates of diseases %	True negative rates of controls %	Accuracy %
kNN with Euclidean metric and equal weighting, $k = 1$	89.1	74.4	83.8
kNN with Euclidean metric and inverse weighting, $k = 1$	89.1	74.4	84.1
kNN with Euclidean metric and squared inverse weighting, $k = 5$	90.1	73.4	**84.5**
kNN with Mahalanobis metric and equal weighting, $k = 1$	90.9	71.4	84.3
kNN with Mahalanobis metric and inverse weighting, $k = 1$	90.9	71.4	84.3
kNN with Mahalanobis metric and squared inverse weighting, $k = 1$	90.9	71.4	84.3
kNN with standardized Euclidean metric and equal weighting. $k = 1$	89.1	74.4	84.1
kNN with standardized Euclidean metric and inverse weighting, $k = 1$	89.1	74.4	84.1
kNN with standardized Euclidean metric and squared inverse weighting, $k = 5$	89.8	73.4	84.3
kNN with Spearman measure and equal weighting, $k = 1$	88.6	66.3	81.1
kNN with Spearman measure and inverse weighting, $k = 5$	89.6	69.3	82.8
kNN with Spearman measure and squared inverse weighting, $k = 5$	90.4	69.3	83.3

Table 4
Results of discriminant analysis, decision tree, multinomial logistic regression, naïve Bayes, random forest and least square (LS) support vector machines (SVM).

Classification method	True positive rates of diseases %	True positive rates of controls %	Accuracy %
Linear discriminant analysis	78.4	62.3	73.0
Mahalanobis discriminant analysis	33.0	94.0	53.5
Quadratic discriminant analysis	76.1	59.3	70.5
Decision tree	89.1	74.4	84.1
Multinomial logistic regression	77.7	62.8	72.7
Naïve Bayes with normal distribution	71.8	67.8	70.5
Naïve Bayes with normal kernel	68.0	78.4	71.5
Naïve Bayes with box kernel	66.5	79.4	70.8
Naïve Bayes with Epanechnikov kernel	67.5	78.9	71.3
Naïve Bayes with triangle kernel	68.8	78.9	72.2
Random forest, number of trees 18	92.4	77.4	**87.4**
LS-SVM with linear kernel, $C = 2^{-4}$	69.3	76.4	71.7
LS-SVM with quadratic kernel, $C = 1$	74.6	77.9	75.7
LS-SVM with cubic kernel, $C = 2^{-5}$	79.4	78.4	79.1
LS-SVM with RBF kernel, $C = 2^{11}$, $\sigma = 2$	91.6	70.9	**84.7**

and make it possible to focus preventive actions on those, and relieve the disease burden from those without any signs of disease at the cellular level.

The high accuracy obtained with our best machine learning algorithm suggests that iPSC technology combined with machine learning could be used even for diagnostic purposes in the future. We will continue to collect more Ca^{2+} transient signals of CMs derived from a larger collection of iPS cell lines carrying different mutations and patient populations, as well as signals of healthy controls, and also from isogenic lines, and thus to improve these methods to make them more suitable for clinical purposes.

Conflicts of interest

None.

Ethical statement

The current study was approved by the Ethics Committee of Pirkanmaa Hospital District, Tampere, Finland, in establishing, culturing and differentiating human iPSC lines (R08070).

Acknowledgments

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imu.2019.01.006.

References

[1] George AL. Molecular and genetic basis of sudden cardiac death. J Clin Invest 2013;123:75–83.
[2] Asatryan B, Medeiros-Domingo A. Translating emerging molecular genetic insights into clinical practice in inherited cardiomyopathies. J Mol Med 2018;96:993–1024.
[3] Kujala K, Paavola J, Lehti A, Larsson K, Pekkarinen-Mattila M, Viitasalo M, Lahtinen AM, Toivonen L, Kontula K, Swan H, Laine M, Silvennoinen O, Aalto-Setälä K. Cell model of catecholaminercig polymorphic ventricular tachycardia reveals early and delayed after depolarizations. PLoS One 2012;7(9)https://doi.org/10.1371/journal.pone.0044660.
[4] Penttinen K, Swan H, Vanninen S, Paavola J, Lahtinen AM, Kontula K, Aalto-Setälä K. Antiarrhythmic effects of Dantrolene in patients with catecholaminergic polymorphic ventricular tachycardia and replication of the responses using iPSC models. PLoS One 2015;10(7)https://doi.org/10.1371/journal.pone.0125366.
[5] Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 2007;131:861–72.
[6] Devalla HD, Passier R. Cardiac differentiation of pluripotent stem cells and implications for modeling the heart in health and disease. Sci Transl Med 2018;10.eaah5457.
[7] Fatima A, Xu G, Shao K, Papadopoulos S, Lehmann M, Arnaiz-Cot JJ, Rosa AO, Nguemo F, Matzkies M, Dittmann S, Stone SL, Linke M, Zechner U, Beyer V, Hennies HC, Rosenkranz S, Klauke B, Parwani AS, Haverkamp W, Pfitzer G, Farr M, Cleemann L, Morad M, Milting H, Hescheler J, Saric T. In vitro modeling of ryanodine receptor 2 dysfunction using human induced pluripotent stem cells. Cell Physiol Biochem 2011;28:579–92.
[8] Jung CB, Moretti A, Mederos y Schnitzler M, Iop L, Storch U, Bellin M, Dorn T, Ruppenthal S, Pfeiffer S, Goedel A, Dirschinger RJ, Seyfarth M, Lam JT, Sinnecker D, Gudermann T, Lipp P, Laugwitz KL. Dantrolene rescues arrhythmogenic RYR2 defect in a patient-specific stem cell model of catecholaminergic polymorphic ventricular tachycardia. EMBO Mol Med 2012;4:180–91.
[9] Novak A, Barad L, Lorber A, Gherghiceanu M, Reiter I, Eisen B, Eldo L, Itskovitz-Eldor J, Eldar M, Arad M, Binah O. Functional abnormalities in iPSC-derived cardiomyocytes generated from CPVT1 and CPVT2 patients carrying ryanodine or calsequestrin mutations. J Cell Mol Med 2015;19:2006–18.
[10] Itzhaki I, Maizels L, Huber L, Gepstein A, Arbel G, Caspi O, Miller L, Belhassen B, Nof E, Glikson M, Gepstein L. Modeling of catecholaminergic polymorphic ventricular tachycardia with patient-specific human-induced pluripotent stem cells. J Am Coll Cardiol 2012;60:990–1000.
[11] Zhang XH, Haviland S, Wei H, Saric T, Fatima A, Hescheler J, Cleemann L, Morad M. Ca^{2+} signaling in human induced pluripotent stem cell-derived cardiomyocytes (iPS-CM) from normal and catecholaminergic polymorphic ventricular tachycardia (CPVT)-afflicted subjects. Cell Calcium 2013;54:57–70.
[12] Di Pasquale E, Lodola F, Miragoli M, Denegri M, Avelino-Cruz JE, Buonocore M, Nakahama H, Portararo P, Bloise R, Napolitano C, Condorelli G, Priori SG. CaMKII inhibition rectifies arrhythmic phenotype in a patient-specific model of catecholaminergic polymorphic ventricular tachycardia. Cell Death Dis 2013;4:e843.
[13] Moretti A, Bellin M, Welling A, Jung CB, Lam JT, Bott-Flugel L, Dorn T, Goedel A, Hohnke C, Hofmann F, Seyfarth M, Sinnecker D, Schomig A, Laugwitz KL KL. Patient-specific induced pluripotent stem-cell models for long-QT syndrome. N Engl J Med 2010;363:1397–409.
[14] Matsa E, Rajamohan D, Dick E, Young L, Mellor L, Staniforth A, Denning C. Drug evaluation in cardiomyocytes derived from human induced pluripotent stem cells carrying a long QT syndrome type 2 mutation. Eur Heart J 2011;32:952–62.
[15] Lahti AL, Kujala VJ, Chapman H, Koivisto AP, Pekkanen-Mattila M, Kerkela E, Hyttinen J, Kontula K, Swan H, Conklin BR, Yamanaka S, Silvennoinen O, Aalto-Setala K. Model for long QT syndrome type 2 using human iPS cells demonstrates arrhythmogenic characteristics in cell culture. Dis. Model Mech. 2012;5:220–30.
[16] Kiviaho AL, Ahola A, Larsson K, Penttinen K, Swan H, Pekkanen-Mattila M, Venäläinen H, Paavola K, Hyttinen J, Aalto-Setälä K. Distinct electrophysiological and mechanical beating phenotypes of long QT syndrome type 1-specific cardiomyocytes carrying different mutations. IJC Heart & Vasculature 2015;8:9–31.
[17] Han L, Li Y, Tchao J, Kaplan AD, Lin B, Li Y, Mich-Basso J, Lis A, Hassan N, London B, Bett CG, Tobita K, Rasmusson RL, Yang L. Study familial hypertrophic cardiomyopathy using patient-specific induced pluripotent stem cells. Cardiovasc Res 2014;104(2):258–69. https://doi.org/10.1093/cvr/cvu205. Epub 2014 Sep 10.
[18] Lan F, Lee AS, Liang P, Sanchez-Freire V, Nguyen PK, Wang L, Han L, Yen M, Wang Y, Sun N, Abilez OJ, Hu S, Ebert AD, Navarrete EG, Simmons CS, Wheeler M, Pruitt B, Lewis R, Yamaguchi Y, Ashley EA, Bers DM, Robbins RC, Longaker MT, Wu JC. Abnormal calcium handling properties underlie familial hypertrophic cardiomyopathy pathology in patient-specific induced pluripotent stem cells. Cell Stem Cell 2013;12:101–13.
[19] Ojala M, Prajapati C, Pölönen RP, Rajala K, Pekkanen-Mattila M, Rasku J, Larsson K, Aalto-Setälä K. Mutation-specific phenotypes in hiPSC-derived cardiomyocytes carrying either myosin-binding protein C or α-tropomyosin mutation for hypertrophic cardiomyopathy. Stem Cell Int 2016:1684792https://www.hindawi.com/journals/sci/2016/1684792/.
[20] Lee EK, Tran DD, Keung W, Chan P, Wong G, Chan CW, Costa KD, Li RA, Khine M. Machine learning of human pluripotent stem cell-derived engineered cardiac tissue contractility for automated drug classification. Stem Cell Rep 2017;9:1560–72.
[21] Juhola M, Joutsijoki H, Varpa K, Saarikoski J, Rasku J, Iltanen K, Laurikkala J, Hyyrö H, Ávalos-Salguero J, Siirtola H, Penttinen K, Aalto-Setälä K. On computation

of calcium cycling anomalies in cardiomyocytes data. 36th Annual Int. Conf. IEEE Eng. Med. Biol. Society. 2014. p. 1444–7. Chicago, Illinois, USA.
[22] Juhola M, Penttinen K, Joutsijoki H, Varpa K, Saarikoski J, Rasku J, Siirtola H, Iltanen K, Laurikkala J, Hyyrö H, Hyttinen J, Aalto-Setälä K. Signal analysis and classification methods for calcium transient data of stem cell derived cardiomyocytes. Comput Biol Med 2015;61:1–7.
[23] Juhola M, Joutsijoki H, Penttinen K, Aalto-Setälä K. Detection of genetic cardiac diseases by Ca^{2+} transient profiles using machine learning methods. Sci Rep 2018;8:9355https://doi.org/10.1038/s41598-018-27695-5.
[24] Han J, Kamber M, Pei J. Data mining: concepts and techniques. third ed. Morgan Kaufmann; 2012.
[25] Duda RO, Hart PE, Stork DG. Pattern classification. second ed. John Wiley & Sons; 2001.
[26] Wu X, Kumar V, Quinlan R, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D. Top 10 algorithms in data mining. Knowl Inf Syst 2008;14(1):1–37.
[27] Tharwat A, Tarek G, Abdelhameed I, Aboul Ella H. Linear discriminant analysis: a detailed tutorial. AI Commun 2017;30(2):169–90.
[28] Cios KJ, Pedrycz W, Swiniarski RW, Kurgan LA. Data mining: a knowledge discovery approach. Springer; 2007.
[29] Wu W, Mallet Y, Walczak B, Penninckx W, Massart DL, Heuerding S, Erni F. Comparison of regularized discriminant analysis, linear discriminant analysis and quadratic discriminant analysis, applied to NIR data. Anal Chim Acta 1996;329(3):257–65.
[30] G. Bohling, Classical normal-based discriminant analysis, Technical report, Kansas Geol Surv. http://people.ku.edu/gbohling/EECS833 Accessed 26.7.2018.
[31] Hastie T, Tibshirani R, Friedmann J. The elements of statistical learning – data mining, inference, and prediction. second ed. Springer; 2009.
[32] Ng AY, Jordan MI. On discriminative vs. generative classifiers: a comparison of logistic regression and naïve Bayes NIPS. 2002. p. 841–8.
[33] Agresti A. Categorical data analysis. John Wiley & Sons; 1990.
[34] Kwak C, Clayton-Matthews A. Multinomial logistic regression. Nurs Res 2002;51(6):404–10.
[35] Rokach L, Maimon O. Data mining with decision trees: theory and applications. World Scientific; 2015.
[36] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
[37] Chen W, Xie X, Wang J, Pradhan B, Hong H, Bui DT, Duan Z, Ma J. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. Catena 2017;151:141–60.
[38] Couronné R, Probst P, Boulesteix A-L. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinf 2018;19:270.
[39] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97.
[40] Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J. Least squares support vector machines. World Scientific; 1999.
[41] Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. Neural Process Lett 1999;9(3):293–300.
[42] Van Gestel T, Suykens JAK, Baesens B, Viaene S, Vanthienen J, Dedene G, De Moor B, Vandewalle J. Benchmarking least squares support vector machine classifiers. Mach Learn 2004;54(1):5–32.
[43] Casini S, Verkerk AO, Remme CA. Human iPSC-derived cardiomyocytes for investigation of disease mechanism and therapeutic strategies in inherited arrhythmia syndromes: strengths and limitations. Cardiovasc Drugs Ther 2017;31:325–44.
[44] Veerman CC, Kosmidis G, Mummery CL, Casini S, Verkerk AO, Bellin M. Inmaturity of human stem-cell-derived cardiomyocytes in culture: fatal flaw or soluble problem. Stem Cell Dev 2015;24(9):1035–52. https://doi.org/10.1089/scd.2014.0533.
[45] Wilde AA, Behr ER. Genetic testing for inherited cardiac disease. Nat Rev Cardiol 2013;10:571–83.