Computers in Biology and Medicine 142 (2022) 105218

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Data analytics for cardiac diseases

Martti Juhola [a,*], Henry Joutsijoki [a], Kirsi Penttinen [b], Disheet Shah [c], Risto-Pekka Pölönen [d], Katriina Aalto-Setälä [b,e]

[a] Faculty of Information Technology and Communication Sciences, Tampere University, 33014, Tampere, Finland
[b] Faculty of Medicine and Health Technology, Tampere University, 33014, Tampere, Finland
[c] Department of Pharmacology, Northwestern University, Chicago, IL, 60611, USA
[d] Department of Pharmacology, University of California Davis, 95616, Davis, CA, USA
[e] Heart Center, Tampere University Hospital, 33520, Tampere, Finland

ARTICLE INFO

Keywords:
Data analytics
Peak detection
Calcium transient signals
Cardiac diseases

ABSTRACT

In the present research we tackled the classification of seven genetic cardiac diseases and control subjects by using an extensive set of machine learning algorithms with their variations from simple K-nearest neighbor searching method to support vector machines. The research was based on calcium transient signals measured from induced pluripotent stem cell-derived cardiomyocytes. All in all, 55 different machine learning alternatives were used to model eight classes by applying the principle of 10-fold crossvalidation with the peak data of 1626 signals. The best classification accuracy of approximately 69% was given by random forests, which can be seen high enough here to show machine learning to be potential for the differentiation of the eight disease classes.

1. Introduction

The function of calcium cycling is essential in excitation-contraction coupling of human cardiomyocytes. Abnormal calcium cycling is connected to arrhythmia associated with cardiac disorders and heart failure. The detection and characterization of these distortions are important, since patient phenotype can be identified, and recognition and diagnostics of cardiac diseases can be improved. In this context induced pluripotent stem cell-derived cardiomyocytes (iPSC-CM) are important for the research of genetic cardiac diseases [1].

Patient-specific iPSC-derived cardiomyocytes (iPSC-CMs) offer an attractive experimental platform to model cardiac functionality and diseases. Many different genetic cardiac diseases have been studied with iPSC-technology including long QT syndrome 1 and 2 (LQT1 and LQT2) [2–4], electric disorders of the heart that predisposes patients to arrhythmias and sudden cardiac death [5], dilated cardiomyopathy (DCM) [6], a disease of the heart muscle, hypertrophic cardiomyopathy (HCM) [7,8], disorder that affects the structure of heart muscle tissue leading to arrhythmias and progressive heart failure, Brugada syndrome (BrS) that predisposes patients to fatal cardiac arrhythmias [9,10], and catecholaminergic polymorphic tachycardia (CPVT), an exercise-induced malignant arrhythmogenic disorder [11–13].

Earlier we found that machine learning computation can be used for the analysis and classification of iPSC-CMs [14–16]. Obviously, machine learning has still been seldom applied to this type of data. At least mechanistic action of drugs in cardiology [17] and electrophysiological effects of chronotropic drugs [5] have been studied. The identification of abnormal calcium transients from iPSC-CM data [18] was made with data analytics. Recently, machine learning was utilized to classify healthy and diseased cardiomyocytes on the basis of contractility profiles of cardiomyocytes [19]. Coronnello and Francipane [20] did a review about iPSC applications where artificial intelligence and machine learning methods are used. In this review, the classification of genetic cardiac diseases is also covered and machine learning methods used in the applications are described.

Originally, we began our data analytics research for iPSC-CM data by separating abnormal from normal calcium signals when normal signals contain regular and quite similar sizes of peaks whereas abnormal signals contain at least partially rather irregularly cycling and sometimes asymmetric peaks as to their left and right sides [14]. Thereafter, we found that calcium transient signals associated with genetic cardiac diseases LQT1, HCM due to a myosin-binding protein C (MYBPC3) mutation (HCMM) and CPVT were possible to distinguish from each other as well as from calcium transient signals of healthy controls (WT) [15] by applying machine learning. Then we extended this research also to cover disease HCM due to an α-tropomyosin (TPM1) mutation

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

https://doi.org/10.1016/j.compbiomed.2022.105218
Received 15 September 2021; Received in revised form 17 December 2021; Accepted 3 January 2022
Available online 5 January 2022

(HCMT) [16].

The approach of using machine learning methods for calcium transient signals derived from iPSC-derived cardiomyocytes is useful, because this makes simpler and easier to analyze and separate rapidly between different genetic cardiac diseases. Thus far, this is made by means of biopsy into a patient's heart, but iPSC cardiomyocytes originate from stem cells cultured and reprogrammed, e.g., from the skin sample of a patient. These iPSC-derived cardiomyocytes are genetically equivalent with a patient's actual cardiomyocytes. If we compare taking a small skin sample to an invasive biopsy, the former approach is safer for patient, cost-effective and reduces the possible time spent in a hospital. Hence, by means of machine learning, we can improve the safety of a patient.

In the present research we expanded our data to cover classes LQT2, DCM and BrS. We also expanded additional signals to the classes of CPVT and WT. Furthermore, we classified the data with 55 classifier types, several more than earlier [15,16]. This study is novel, since, to the best of our knowledge no one has earlier made this type of research by applying data of several cardiac diseases and an extensive number of different machine learning methods. In a new article [19] they classified healthy cardiomyocytes vs. diseased ones with 17 different classifier alternatives: nine different nearest neighbor searching method variations, three Naive Bayes method variations, decision trees, quadratic discriminant analysis and three support vector machine variations. However, they did not use random forests that produced the best results in our present research. The research [19] only consisted of healthy WT and LQT data and, in this way, their binary classification problem was clearly easier and more limited compared with our current research with eight disease classes and 55 classifier alternatives. Further, machine learning modelling for calcium transient signals produced by iPSC-derived cardiomyocytes is so new that, obviously, only a few other articles [5,17–19] are published so far in addition to our research such as [14–16].

2. Materials

The present research was approved by the Ethics Committee of Pirkanmaa Hospital District associated with culturing and differentiation of human iPSC lines (R08070). Patient-specific iPSC lines were established and cultured as previously described [15]. According to Table 1, iPSC lines were generated from two LQT1 and two LQT2 patients (lines 03412 and 03417 were from the same patient and lines 03809 and 03810 as well), two HCMT and two HCMM patients, six CPVT patients (lines 03701 and 03706 were from the same patient and 05603 and 05605 were from the same patient), one BrS patient, two DCM patients, and two healthy control individuals (WT). CPVT patients carried cardiac ryanodine receptor (RyR2) mutations, BrS patient carried SCN5A mutation, HCMM patients carried myosin-binding protein C (MYBPC3) mutations, HCMT patients α-tropomyosin (TPM1) and LQT1 patients potassium voltage-gated channel subfamily Q member1 (KCNQ1) mutation. LQT2 patients carried the human ether-a-go-go-related gene (HERG) mutation and DCM patients were with lamin A and lamin C (LMNA) mutations. WT data were generated from two healthy control individuals. Patient-specific iPSCs were differentiated into spontaneously beating cardiomyocytes and dissociated as single cells for calcium imaging studies, which were conducted in spontaneously beating Fura-2 AM (Invitrogen, Molecular Probes) or Fluo-4 AM (Thermo Fisher Scientific) loaded cardiomyocytes as described earlier [11–13]. Every calcium transient signal was from a video recording from one cardiomyocyte. The background fluorescence (signal from empty coverslip) was subtracted before converting the videos to traces of intensity over time (Fig. 1). This was performed with the background correction function of the calcium measurement software by determining the background value from noncellular region with the help of a background region of interest (ROI) or a fixed value from the background. The traces were then annotated by the experimenting scientist.

Table 1
Data of 23 cell lines used.

Disease	Cell line	Number of signals per disease
LQT1	00118	90
	00208	
HCMM	06108	270
	07801	
CPVT	03701	361
	03706	
	05208	
	05404	
	05503	
	05603	
	05605	
	07001	
WT	04602	331
	04511	
HCMT	02912	149
	13602	
LQT2	03412	138
	03417	
	03809	
	03810	
DCM	12619	69
	12704	
BrS	14004	218

3. Data extracted from calcium transient signals

In our data there were 1626 signals altogether. The numbers of signals per each disease class are shown in Table 1. A signal was seen abnormal if one or more of its peaks were identified to be abnormal. The algorithm developed earlier [14] could have been used for this, but the types, normal or abnormal, of all entire signals were determined by a biotechnological expert so that we saw them as surely annotated as possible. Previously we found that both normal and abnormal signals together could have been classified into different disease classes (and control cases) without separating their normal and abnormal signals by using machine learning when there were four classes [15] or five classes [16]. Thus, this principle was also followed in the current research, where there were more abnormal than normal signals in four disease classes and 37% of signals were abnormal in HCMM, 44% in HCMT and 44% in BrS. In controls (WT) there were only 14% abnormal showing that, in this respect, these signals were mostly different from those of the seven disease classes. There are two example signals for each of LQT1, HCMM and BrS in Fig. 1. However, notice that sizes and shapes of peaks may vary from one signal to another also subject to the same disease class and particularly as to abnormal signals.

To detect peaks the first derivative of the signal was approximated and utilized to find the beginning, maximum and the end of each peak. At the beginning of a peak the first derivative is close to zero, then increases and after the approximate midpoint of the left peak side decreases close to zero towards the top (maximum) and thereafter changes negative while advancing, but finally after the approximate midpoint of the peak right side becomes back to zero at the end of the peak. After having found possible peaks from a signal very small peaks less than 8% of an estimated mean peak amplitude of the entire signal were left out as possible noise. This peak detection was described more in detail previously [14,15]. When peak forms varied and signal lengths from roughly 8 s–47 s, they also contained various numbers of peaks, from 1 to around 120.

After the detection of valid peaks of a calcium transient signal it was essential to extract values of suitable peak attributes from the peaks accepted for the classification of signals into different classes. We computed 14 peak attributes shown in Fig. 2 as follows. The amplitude of the left side of a peak, amplitude of its right side and their durations were computed. Next the approximated maximum of the 1st derivative of the left side, correspondingly the absolute minimum of the right side

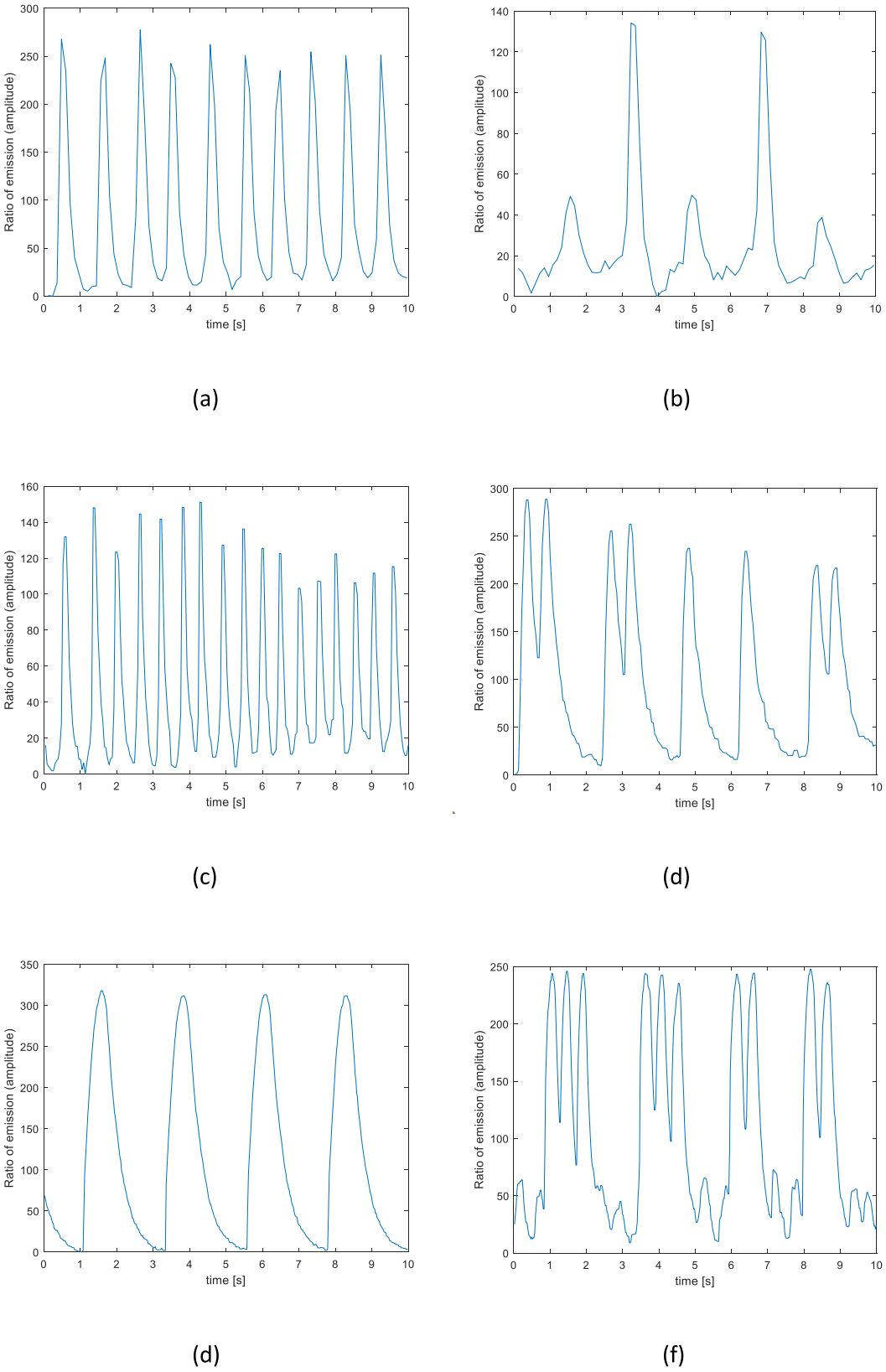

Fig. 1. Example segments of 10 s from (a) a normally cycling LQT1 calcium transient signal, an abnormal LQT1 signal, (c) a normal HCMM signal, (d) an abnormal HCMM signal of oscillating peaks, (e) a normal BrS signal with very regular peak shapes and sizes and (f) an abnormal BrS transient signal containing oscillating and possibly small peaks.

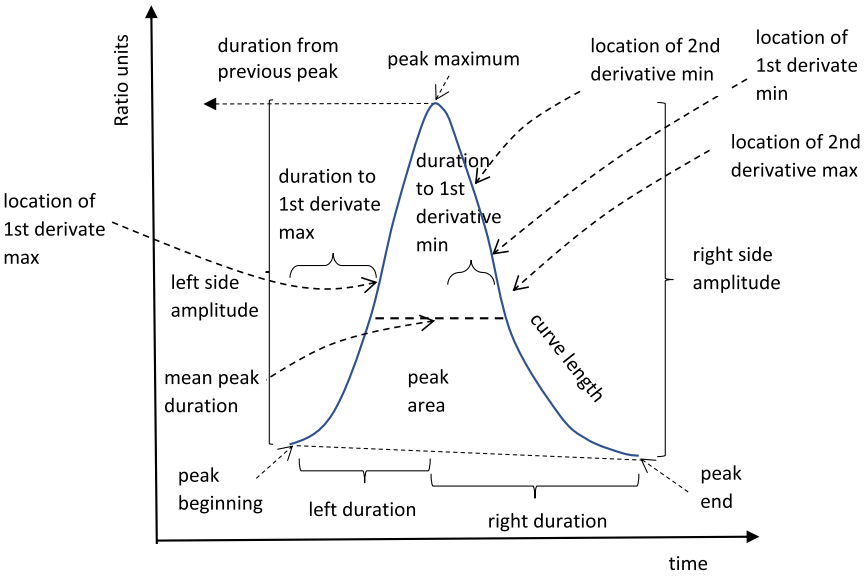

Fig. 2. Peak amplitudes of left side and right side, peak left and right durations, approximate location for the 1st derivative maximum, approximate location for the absolute value of 1st derivative minimum, locations for the absolute value of the 2nd derivative minimum and for the 2nd derivative maximum, peak surface area, duration from the previous peak, duration from the peak beginning to the 1st derivative maximum, duration from the peak maximum to the 1st derivative minimum, mean peak duration, and curve length.

and maximum and absolute minimum of the 2nd derivative from the peak right side were computed. Then the area bounded by the peak curve and the line from the peak beginning to its end, duration from the maximum of the preceding peak to that of the current peak or for the first peak of a signal it from the signal beginning, duration from the peak beginning to the first derivative maximum and the duration from the peak maximum to the first derivative absolute minimum were computed. The middle points of the left and right amplitudes were first computed as well as their average and then between the intersection points of this average line with the peak curve in the left and right peak sides the distance was computed. Ultimately, the approximated peak curve length was computed to be the last attribute. The total of peaks computed was 27956.

4. Technical background for experiments

4.1. Methods

In this paper, a wide collection of machine learning methods was applied to the classification task. We have used the majority of the classification methods of this study also in our earlier studies [14–16,21,22]. This enables a long-term perspective about the suitability and possibilities of each classification method for the classification of genetic cardiac diseases. In other words, we can see the impact when the number of classes (i.e. cardiac diseases) increases and its effect on the classification. However, we have added to this paper previously unused methods to advance the development of machine learning methods in this field.

When different methods are tested or developed for the classification of genetic cardiac diseases, we need to separate scientific and practical point of views between each other. If the classification task is seen purely from the scientific point of view, it requires that the methods used are novel complex deep learning techniques, which can be understood only if a person has a proper technical background and adequate knowledge about machine learning. However, end users who use the software and/or algorithms in practice to identify whether or not a patient has a genetic cardiac disease do not usually have such technical background. End users are the first persons from which a patient will ask what is done to him/her or what the software/algorithm do. Hence, the complexity of algorithm may set some limitations for its use.

Practical perspective instead emphasizes that the method works in practice and does not require that the method is scientifically state-of-the-art level. In practice, if a method is simple and it works, it has several advantages. Firstly, the understandability about the method will be increased also from the end users' side. Secondly, possible productization of the algorithms and processes developed for the classification task is easier and faster when the methods are simple and have stable theoretical basis. This again reduces the costs regarding the productization.

k-Nearest Neighbor classifier (KNN) [23] is a widely used distance-based classification method that has three main components. The first component is to select the *k* value describing the number of closest training examples with respect to the test example in prediction phase. Selecting odd *k* value decreases the possibility of ties in prediction. The second component is distance measure that explains the way, how distances between examples are evaluated. The third component is distance weighting and with the help of it higher weight can be given for the *k* closest training examples. There are not existing rules how to select the optimal parameter values for KNN, but they are empirically examined.

Linear discriminant analysis (LDA) [24], quadratic discriminant analysis (QDA) [25] and Mahalanobis discriminant analysis (MDA) [26] are variants of discriminant analysis. Discriminant analysis methods are statistically oriented and can be used in both binary and multi-class classification tasks. The main idea of discriminant analysis methods is to divide the input space into multiple regions by creating decision boundaries such that each region should represent one class. Decision boundaries can be linear or non-linear depending on the method used. Discriminant analysis methods do not include parameters to be tuned.

Naïve Bayes (NB) is a classification method for both binary and multi-class cases that can be used with and without kernel density estimation (KDE) [27]. NB has the assumption of independent features and when the features are numerical, Gaussian distribution is assumed. Both of these assumptions are strong in real-life. When predicting class for new test example, NB provides a probability for each class and the class having the highest probability will be chosen as a final prediction for the test example. NB can be extended with KDE, which is a method for smoothing the probability density function. There exist different kernels for KDE and in this work four different kernels were tested (see Table 2 for details). Multinomial logistic regression (MLR) [28] is a straightforward extension from logistic regression method designed for binary classification tasks. MLR applies maximum likelihood estimation when constructing the generalized logit model. Overall, MLR requires *N-1* logit equations for *N* class classification problem. Although MLR uses iterative algorithm such as Newton-Raphson method when constructing the classification model, it does not include tunable

Table 2
Parameter value spaces tested for classification methods used in the paper. CART, LDA, MDA, MLR and QDA had none such.

Method	Trees	*k* value	Distance measure	Distance weighting	Kernel (KDE)	Coding design	Box constraint	Kernel scaling σ (RBF)	Kernel (SVM)
ECOC-SVM	–	–	–	–	–	{one-vs.-all (OVA), one-vs.-one (OVO), ordinal, dense random, sparse random}	$\{2^{-8}, 2^{-7}, \ldots, 2^{12}\}$	$\{2^{-8}, 2^{-7}, \ldots, 2^{12}\}$	{linear, quadratic, polynomial 3, RBF}
KNN	–	{1,3,5, …,37}	{Chebysev, cityblock, correlation, cosine, Euclidean, Mahalanobis, standardized Euclidean, Spearman}	{equal, inverse, squared inverse}	–	–	–	–	–
NB	–	–	–	–	{normal, box, Epanech-nikov, triangle}	–	–	–	–
RF	{1,2,3, …,150}	–	–	–	–	–	–	–	–

hyperparameters.

Classification and regression tree (CART) [29] and random forests (RF) [30] are tree-based classification methods. The output of CART algorithm is a decision tree model that can be used for predicting new test examples and consists of inner nodes and leaf nodes. In inner nodes, a variable and a splitting value of it is represented. Based on the variable and the splitting value, a new test example traverses in the tree via inner nodes until a leaf node is encountered where the predicted label for the test example exists. The search of suitable variable and splitting value for each inner node is performed in the training phase of CART algorithm. The size of a decision tree can be adjusted by changing parameters such as tree depth or a minimum number of examples for splitting node. However, in this study we applied the default parameter values when training the CART algorithm.

Random forest (RF) is an extension compared to CART algorithm since it includes several individual decision trees. The way how a decision tree is formed in RF differs from CART algorithm. Firstly, RF uses a random subset from the training set to construct the individual tree model. Secondly, in the case of each inner node, RF selects a random subset from available features and searches for the best splitting feature within the selected feature set. The most important parameter in RF is to select how many trees are in a forest and Table 2 shows what parameter values were tested for the number of trees in our research.

Support Vector Machine (SVM) [31] is a classification method that aims at finding a classes separating hyperplane such that the margin between classes is maximized. The first version of SVM was designed for linearly separable cases, but later the use of SVM was extended to linearly nonseparable cases with the kernel trick. In kernel trick, the data are transformed to a higher dimensional space where a classes separating hyperplane can be found. There are different kernels for SVM and in Table 2 kernels used in this study can be found. Depending on the kernel choice, SVM includes various parameters. A common parameter for SVM and, thus, for all kernels is boxconstraint that controls the trade-off between maximizing the margin and the number of misclassified points. For the Radial Basis Function (RBF) kernel, there is also a parameter that specifies the width of Gaussian function. We have tested different combinations of boxconstraint and kernelscale for the RBF function and from Table 2 the parameter value spaces can be seen.

SVM is a two-class classification method and we have a multi-class classification task. Hence, we have applied error-correcting output codes (ECOC) framework [32–34] to model the multi-class SVM classification task. Moreover, we have used ISDA [35] optimization algorithm to solve the optimization task for a binary SVM classifier. ECOC contains two phases called encoding and decoding. Encoding phase contains deciding coding matrix where the columns depict binary SVM classifiers and rows represent codewords for classes. Depending on the coding design, coding matrix can be constructed using binary or ternary coding. In this study, we have used five different coding designs: one-vs.-all (OVA), one-vs.-one, ordinal, dense random and sparse random. In OVA, each class is separated from the rest with one SVM classifier whereas in OVO an SVM classifier is constructed for each pair of classes. Ordinal (ORD) coding design is constructed such that the first SVM classifier separates the first class from the rest. The second classifier separates the first two classes from the rest etc. In dense random coding design classes are assigned randomly to +1/-1 label so that both classes are represented at least once. The last coding design, sparse random, is a modification compared to dense random where classes are assigned +1/-1 label or ignored with a given probability. When coding design is constructed and binary SVM classifiers are trained based on the coding matrix, each SVM classifier gives a prediction for the test example. By collecting the predictions together, a codeword is created for the test example. In decoding phase, the closest or loss function minimizing codeword from coding matrix with respect to test example's codeword assigns the final prediction.

4.2. Classification settings

Classification was performed using peak-based dataset where the columns represent the attributes extracted from each peak and they are explained in Section 3 in detail. Each row presents the data from a single peak. Classification with all classification methods presented in Section 4.1 was performed using 10-fold cross-validation on a signal level such that the whole data from specific signal is included only to one fold. Cross-validation division was made using stratified sampling so that in each fold all classes are represented. The same cross-validation division was used with all machine learning methods and parameter settings tested in order to have a fair comparison between the results of the classification methods. Furthermore, in the case of ECOC, dense and sparse random coding designs include randomness and, thus, we fixed the coding matrix for each training set in cross-validation procedure when examining different parameter settings. When performing the cross-validation, in each cross-validation round the training set was z-score standardized to have a zero mean and unit variance. Then the data in test set were scaled based on the scaling parameters obtained from a training set.

Training of classification methods was performed using peak-level data whereas the results presented are signal-level results. A classification method gives a prediction for each peak within a signal. In order to have a signal-level prediction, we take a mode from the peak-level predictions within a signal. Accuracy and sensitivity were selected as the performance measures to this study and these performance measures were evaluated from signal level predictions in percentages and from

each cross-validation test set separately. Since we used 10-fold cross-validation, the mean and standard deviation were calculated from the performance measures. Table 2 shows the parameter value spaces tested for the methods described in Section 4.1. Parameter tuning was made by applying grid search. For example, ECOC SVM with OVA coding design and the RBF kernel 21^2 combinations of boxconstraint and kernel scale parameter values were tested. Cross-validation was repeated with all parameter combinations tested and the highest accuracy determined the best parameter value/values. Otherwise, we have used default parameter settings. Experiments were performed using Matlab 2019b, Statistics and Machine Learning Toolbox and Parallel Computing Toolbox.

5. Results

Classification results were computed in several ways by applying the above machine learning methods. Table 3 shows the results of random forests and various constructions of support vector machine methods. Table 3 also presents the results of discriminant analysis, Naïve Bayes, decision trees constructed with CART method and multinomial logistic regression. Random forests as an ensemble method produced the best classification accuracy. Support vector machines with a radial basis function (RBF) were the next best and then CART. Table 4 describes the results of several nearest neighbor searching methods used with different distance measures. Nearest neighbor searching method with Mahalanobis inverse weighting generated the best accuracy and that with Mahalanobis squared inverse weighting the next best. When random forests in Table 3 gave the best classification accuracy of all, its confusion matrix is shown in Table 5. The diagonal showing true positive rates (sensitivities) of the confusion matrix indicates that LQT1, HCMM and CPVT classes can be classified well, but HCMT and DCM classes are somewhat more complex and WT, LQT2 and BrS more complex to classify. Further, we notice that HCMT is intermingled most with HCMM, since the false negative percent of HCMT is as high as 24.8%. Correspondingly, LQT2, DCM and BrS caused their highest false negative percents in the column of WT.

6. Discussion

By applying several machine learning methods and their variation we modeled genetic cardiac diseases and controls (WT) based on calcium transient signals of IPSC-CMs. Modeling here means building computationally classifiers that are capable to distinguish test signals from among these eight classes. Classes LQT1 and HCMM could be separated best and CPVT well. Furthermore, HCMT and DCM were next with classification accuracies greater than 53%, but WT, LQT2, and BrS were the poorest with accuracies greater than 45%. Rather low sensitivities or true positive rates show that the data is complex and data items of some classes resemble somehow those of other classes causing false negative outcomes. Nevertheless, the results obtained are reasonable and indicate that the separation of even eight different classes of the current data is promising, because in each row of Table 5 the correct class on the diagonal received the clearly highest percent even if the classification accuracy of all data was 69%. This is the major result. Notice that it is very relative which is a very good or good result for various machine learning classification. Rarely classification accuracies of over 90% are gained or even 80%. The properties of a data set affect highly. Particularly, the number of classes is essential. Usually, it is much easier and probable to achieve good results if the number of classes is equal to the minimum of two called binary classification. Thus, the greater number of classes, the more complex classification is encountered. In our previous studies the highest classification accuracy of 78.6% was attained for 527 signals of four classes LQT1, HCMM, CPVT and WT and the accuracy of 87.6% when only three disease classes were classified without WT [15]. Later the classification accuracy of

Table 3
Classification results in percent as average sensitivities (true positive rates) of the classes and their average accuracies also with standard deviations given by random forests, support vector machines (SVM) with hyperparameters C and σ, discriminant analysis, Naïve Bayes, CART and multinomial logistic regression. The four best accuracies are written in Bold given by random forests, two support vector machine classifiers with radial basis function (RBF), and decision tree method called CART.

Method	Sensitivity %								Accuracy %
Class	LQT1	HCMM	CPVT	WT	HCMT	LQT2	DCM	BrS	
Random forests, 100 trees	90.0	85.9	70.3	61.0	60.4	54.5	58.3	66.5	**68.8 ± 4.1**
ECOC-OVO-SVM linear, $C = 2^0$	77.8	64.4	39.9	48.3	46.4	30.4	47.6	38.5	47.7 ± 4.4
ECOC-OVO-SVM quadratic, $C = 2^{-4}$	87.8	81.1	55.7	45.9	55.0	46.3	53.8	48.1	57.8 ± 2.4
ECOC-OVO-SVM polynomial 3, $C = 2^{-8}$	93.3	82.6	57.3	49.9	53.8	58.8	59.8	24.3	57.4 ± 4.0
ECOC-OVO-SVM RBF, $C = 2^9$, $\sigma = 2^2$	86.7	84.8	64.2	53.5	64.4	53.6	53.8	59.2	**64.7 ± 4.1**
ECOC-OVA-SVM linear, $C = 2^{-2}$	97.8	44.8	6.6	17.8	23.3	26.0	39.3	54.7	31.3 ± 2.0
ECOC-OVA-SVM quadratic, $C = 2^{-6}$	98.9	83.7	42.3	32.0	41.5	46.9	65.5	61.9	54.1 ± 5.3
ECOC-OVA-SVM polynomial 3, $C = 2^{-8}$	80.0	54.8	32.1	42.0	40.4	46.2	43.8	30.7	42.8 ± 5.4
ECOC-OVA-SVM RBF, $C = 2^8$, $\sigma = 2^2$	91.1	83.0	59.2	51.7	68.4	52.9	46.4	54.6	62.5 ± 4.1
ECOC-Ordinal-SVM linear, $C = 2^{-1}$	0.0	0.0	41.0	55.6	1.3	39.1	0.0	11.0	25.3 ± 3.9
ECOC-Ordinal-SVM quadratic, $C = 2^{-6}$	62.2	47.0	38.8	61.0	14.8	60.7	30.2	17.0	42.4 ± 4.0
ECOC-Ordinal-SVM polynomial 3, $C = 2^{-8}$	83.3	56.3	29.6	33.3	10.1	16.9	21.4	35.3	35.3 ± 2.5
ECOC-Ordinal-SVM RBF, $C = 2^5$, $\sigma = 2^1$	83.3	74.4	59.8	59.8	55.7	58.0	49.2	43.1	60.3 ± 3.7
ECOC- dense random -SVM linear, $C = 2^1$	2.2	63.0	38.3	46.2	2.7	10.2	18.6	42.7	36.1 ± 3.4
ECOC- dense random -SVM quadratic, $C = 2^{-6}$	86.7	75.2	44.0	48.9	25.5	39.2	43.8	46.8	50.9 ± 3.5
ECOC- dense random -SVM polynomial 3, $C = 2^{-8}$	72.2	24.4	39.0	35.0	33.1	33.2	46.7	41.7	37.3 ± 5.3
ECOC-dense random-SVM RBF, $C = 2^8$, $\sigma = 2^1$	84.4	84.4	64.8	52.0	62.4	46.4	45.0	52.7	62.3 ± 3.8
ECOC- sparse random -SVM linear, $C = 2^1$	52.2	65.2	30.4	46.2	15.3	14.0	33.3	43.6	39.7 ± 3.5
ECOC- sparse random -SVM quadratic, $C = 2^{-5}$	90.0	76.3	46.0	49.5	46.3	41.4	43.8	46.3	53.8 ± 4.4
ECOC- sparse random -SVM polynomial 3, $C = 2^{-8}$	84.4	72.2	40.4	38.6	48.4	42.3	58.3	34.7	48.7 ± 5.0
ECOC- sparse random -SVM RBF, $C = 2^9$, $\sigma = 2^2$	85.6	83.3	64.8	52.0	66.4	47.1	50.7	61.9	**64.1 ± 3.5**
Linear discriminant analysis	40.0	54.1	39.3	47.7	16.8	18.0	51.2	33.9	39.4 ± 2.1
Quadratic discriminant analysis	86.7	76.7	23.5	41.1	47.0	18.1	74.0	23.4	43.2 ± 4.6
Mahalanobis discriminant analysis	7.8	7.4	39.0	48.3	4.7	37.7	66.7	54.5	33.9 ± 3.2
Naïve Bayes	60.0	60.7	7.8	23.9	9.3	12.3	81.2	22.0	28.3 ± 2.7
Naïve Bayes normal kernel density estimation	52.2	63.3	8.0	40.5	31.6	20.9	68.3	21.1	33.8 ± 4.0
Naïve Bayes box kernel density estimation	46.7	61.2	11.6	42.6	22.9	16.6	71.4	22.5	33.6 ± 3.6
Naïve Bayes Epanechnikov kernel density estimation	48.9	59.6	8.9	42.0	28.9	15.8	68.3	23.0	33.1 ± 3.4
Naïve Bayes triangle density estimation	53.3	63.7	8.9	40.8	31.0	21.6	68.3	23.0	34.4 ± 3.8
CART (decision tree)	87.8	83.7	68.7	52.0	54.4	50.1	55.5	52.3	**63.2 ± 3.5**
Multinomial logistic regression	85.6	55.9	31.8	41.4	29.5	31.8	54.8	38.6	42.4 ± 4.3

Table 4
Classification results in percent as average sensitivities (true positive rates) of the classes and their average accuracies produced by nearest neighbor searching methods (KNN) with varying numbers k of nearest neighbors calculated. The two best accuracies are written in Bold generated by KNN with Mahalanobis distance measure.

Method	Sensitivity %								Accuracy %
Class	LQT1	HCMM	CPVT	WT	HCMT	LQT2	DCM	BrS	
KNN Chebysev equal weighting, $k = 1$	70.0	83.7	52.9	46.2	50.4	49.3	33.3	46.8	55.4 ± 5.6
KNN Chebysev inverse weighting, $k = 5$	66.7	84.8	47.6	55.6	51.7	46.3	42.4	45.9	56.3 ± 4.6
KNN Chebysev squared inverse weighting, $k = 5$	68.9	84.4	49.3	54.7	53.0	47.0	40.7	45.9	56.6 ± 4.7
KNN cityblock equal weighting, $k = 3$	87.8	86.3	53.4	53.5	57.8	52.3	48.1	49.6	60.3 ± 3.6
KNN cityblock inverse weighting, $k = 3$	85.6	86.7	54.8	55.6	57.8	52.9	48.1	49.5	61.1 ± 4.0
KNN cityblock squared inverse weighting, $k = 3$	84.4	87.0	54.8	55.0	58.4	53.7	48.1	50.0	61.1 ± 4.3
KNN correlation equal weighting, $k = 3$	82.2	73.0	55.9	52.0	55.0	51.4	56.7	51.8	58.4 ± 5.0
KNN correlation inverse weighting, $k = 3$	80.0	77.4	57.3	51.4	57.0	50.6	52.4	49.0	58.8 ± 5.0
KNN correlation squared inverse weighting, $k = 3$	82.2	78.5	56.8	49.6	59.1	50.6	52.4	50.9	59.1 ± 5.7
KNN cosine equal weighting, $k = 1$	78.9	78.1	54.8	55.6	56.4	50.7	51.0	50.9	59.3 ± 4.7
KNN cosine inverse weighting, $k = 1$	78.9	78.1	54.8	55.6	56.4	50.7	51.0	50.9	59.3 ± 4.7
KNN cosine squared inverse weighting, $k = 1$	78.9	78.1	54.8	55.6	56.4	50.7	51.0	50.9	59.3 ± 4.7
KNN Euclidean equal weighting, $k = 1$	77.8	83.7	52.6	48.9	56.4	49.1	46.7	47.2	57.5 ± 2.9
KNN Euclidean inverse weighting, $k = 5$	77.8	81.9	52.0	54.1	57.0	47.8	46.9	48.6	58.2 ± 3.1
KNN Euclidean squared inverse weighting, $k = 5$	78.9	82.2	53.4	53.8	57.0	47.1	48.3	48.2	58.6 ± 2.9
KNN Mahalanobis equal weighting, $k = 1$	82.2	85.9	58.7	56.2	59.8	55.9	46.2	50.4	62.2 ± 3.4
KNN Mahalanobis inverse weighting, $k = 3$	87.8	85.9	59.5	55.0	57.0	59.5	55.2	53.1	**66.3 ± 4.4**
KNN Mahalanobis squared inverse weighting, $k = 3$	86.7	85.9	60.1	54.7	57.0	59.5	55.2	53.1	**63.3 ± 4.2**
KNN standardized Euclidean equal weighting, $k = 1$	77.8	83.7	52.6	48.9	56.4	49.1	46.7	47.2	57.5 ± 2.9
KNN standardized Euclidean inverse weighting, $k = 5$	77.8	81.9	52.0	54.1	57.0	47.8	46.9	48.6	58.2 ± 3.2
KNN standardized Euclidean squared inverse weighting, $k = 5$	78.9	82.2	53.4	53.7	57.0	47.1	48.3	48.2	58.6 ± 2.9
KNN Spearman equal weighting, $k = 1$	85.6	73.0	53.4	53.5	56.4	41.9	43.6	50.9	57.0 ± 2.2
KNN Spearman inverse weighting, $k = 3$	84.4	73.3	55.9	54.4	54.3	42.6	55.0	52.3	58.3 ± 4.1
KNN Spearman squared inverse weighting, $k = 3$	84.4	73.7	57.3	54.1	57.0	43.3	53.6	50.9	58.7 ± 3.3

Table 5
Confusion matrix with the percent results of random forests in the rows. The diagonal contains sensitivities or correctly classified signals class by class. Otherwise, the rows show false negatives row by row for the classes.

True classes	Predicted classes							
	LQT1	HCMM	CPVT	WT	HCMT	LQT2	DCM	BrS
LQT1	**90**	0	8.9	0	0	1.1	0	0
HCMM	0	**85.9**	1.5	6.7	3.7	0.7	0	1.5
CPVT	6.4	2.5	**70.3**	9.2	2.2	2.5	1.4	5.6
WT	2.7	10.8	14.8	**45.9**	4.8	6.6	2.7	11.5
HCMT	0	24.8	4.7	15.5	**55**	0	0	0
LQT2	0	0.8	7.4	23.8	0.7	**46.3**	2.3	18.7
DCM	0	0	10.2	24.5	0	4.3	**53.8**	7.1
BrS	0	2.3	9.1	25.3	0	14.3	0.9	**48.1**

77.8% was gained for 941 signals after having added the fifth class HCMT and some new WT signals [16]. In the current data there were also LQT2, DCM and BrS classes and more signals in CPVT and WT than earlier [15,16].

Obviously, the peaks (their attribute values) of LQT1, HCMM and CPVT differ clearly most from those of all other classes. Conversely, peaks of WT, LQT2 and BrS differ less from some other classes. These are seen indirectly from the rows in Table 5, where there are true positive rates (sensitivities) in per cent on the diagonal and false negative rates in per cent row by row. For LQT1, HCMM and CPVT the false negative rates are small and the largest value of these is 9.2% of CPVT as classified incorrectly to be from class WT. While counting false negative rates greater than 10%, for WT 10.8% of signals were incorrectly classified to be as if from HCMM, 14.8% from CPVT and 11.5% from BrS. Correspondingly for LQT2, 23.8% of signals were incorrectly classified to be from WT and 18.7% from BrS. For BrS 25.3% of signals were classified incorrectly to be from WT and 14.3% from LQT2. On one hand, WT is difficult to separate from some other and, on the other hand, if we look at the column of WT in Table 5 all except LQT1 is somewhat mixed with WT.

Sensitivity percent values less than 50% on the diagonal of the confusion matrix in Table 5 are also important for the current data since, after all, they are the maxima of those rows. In addition, the character of this data is specific when the classification computation was based on individual peaks the majority of which in a signal predicted a class label of that signal. When the sensitivity per cent (on the diagonal) of every disease or control class was clearly the greatest one of every row in Table 5 and there was also a relatively good classification accuracy of 69%, this is the essential positive finding of our study. Still, pondering on practical applications of iPSC-CM data, this process described is not the last objective, but it will be to predict or classify dozens of signals originating from cultured cells of the same subject and then see which class is the most probable in the form of majority for these signals. Therefore, to reach the maximum for the correct disease class or controls will be the principal and most important objective. However, this last test was not yet performed in the current study, because the data were still only from 23 cell lines or at the same time from 19 subjects. Naturally, more data, calcium transient signals from more patients, are needed in future research to ensure our promising finding. From methodological point of view, the next possible step would be to analyze the ensemble approach more detailed way in classification of genetic cardiac diseases context. Although the random forest is an ensemble method and used already in this paper, in future combining of several machine learning methods to an ensemble or several machine learning models from one classification method will be interesting to examine. This paper works as a preliminary study for the in-depth examination of ensemble approach, since constructing the ensemble requires the knowledge of how individual classification methods succeed in

classifying genetic cardiac disease classes, and to this question our study answers. Another approach is to ensemble auto-encoders [36] to cardiac disease classification and in this approach for each class the auto-encoder architecture can be tailored and take into account the possible limitations of available training data.

7. Summary

By studying a complex classification task we gained the classification accuracy of approximately 69%. In our previous studies we have anticipated that genetic cardiac diseases could be separated from each other by machine learning techniques. Here our results showed that separation of even eight different classes of diseases containing controls is possible in sufficient extent. We classified on the basis of individual transient signals, but not yet on the basis of patients (or cell lines), since the number of these was still small. Nevertheless, when even a few hundred signals were derived from cell lines of each class, it is the most important result that their clear majority represents the correct class. The results in Table 4 indicated this to be true, because each correct class (row) was represented by the great majority on the diagonal (sensitivity) of Table 4. To conclude, in this way, the classification accuracy of 69% of random forests is high enough to separate the eight class well enough.

In the future to gain even higher classification accuracies larger datasets collected from several diseases, patients and cell lines would be needed. Also, culture and assay methods should be standardized when aiming to develop predictive computational models [37]. Improvement of the maturation state of the iPSC-CMs, which is common interest in genetic cardiac disease modeling, could enhance the quality of the CMs. Therefore, their calcium handling properties would be more mature and that way diminish the variability of the data sets thus improving the classification accuracy. iPSC cardiac disease modelling enables the study of disease pathophysiology and the development of therapies but also, as shown in this study, it can offer a tool for disease diagnostics. In the future, when CM culturing techniques are improved this machine learning classification method approach could be exploited to diagnose genetic cardiac disease and implemented to evaluate arrhythmia risks of an individual.

Funding

We would like to thank Academy of Finland Centre of Excellence in Body-on-Chip Research.

Declaration of competing interest

None declared.

References

[1] K. Takahashi, K. Tanabe, M. Ohnuki, M. Narita, T. Ichisaka, K. Tomoda, et al., Induction of pluripotent stem cells from adult human fibroblasts by defined factors, Cell 131 (2007) 861–872.
[2] A.L. Kiviaho, A. Ahola, K. Larsson, K. Penttinen, H. Swan, M. Pekkanen-Mattila, H. Venäläinen, K. Paavola, J. Hyttinen, K. Aalto-Setälä, Distinct electrophysiological and mechanical beating phenotypes of long QT syndrome type 1-specific cardiomyocytes carrying different mutations, Int. J. Cardiol. Heart Vasc. 25 (8) (2015) 19–31.
[3] J. Kuusela, K. Larsson, D. Shah, et al., Low extracellular potassium prolongs repolarization and evokes early after depolarization in human induced pluripotent stem cell-derived cardiomyocytes, Biol. Open 6 (2017) 777–784.
[4] D. Shah, C. Prajapati, K. Penttinen, R.M. Cherian, J.T. Koivumäki, A. Alexanova, J. Hyttinen, K. Aalto-Setälä, hiPSC-derived cardiomyocyte model of LQT2 syndrome derived from asymptomatic and symptomatic mutation carriers reproduces clinical differences in aggregates but not in single cells, Cells 9 (5) (2020) 1153, 7.
[5] H. Hwang, R. Liu, J.T. Maxwell, J. Yang, C. Xu, Machine learning identifies abnormal Ca^{2+} transients in human induced pluripotent stem cell-derived cardiomyocytes, Sci. Rep. 10 (2020), 16977.
[6] D. Shah, L. Virtanen, C. Prajapati, M. Kiamehr, J. Gullmets, G. West, J. Kreutzer, M. Pekkanen-Mattila, T. Heliö, P. Kallio, P. Taimen, K. Aalto-Setälä, Modeling of LMNA-related dilated cardiomyopathy using human induced pluripotent stem cells, Cells 8 (6) (2019) 594.
[7] 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 Cell. Int. (2016), 1684792.
[8] C. Prajapati, M. Ojala, K. Aalto-Setälä, Divergent effects of adrenaline in human induced pluripotent stem cell-derived cardiomyocytes obtained from hypertrophic cardiomyopathy, Dis. Model Mech. 11 (2018), dmm032896.
[9] P. Liang, K. Sallam, H. Wu, et al., Patient-specific and genome-edited induced pluripotent stem cell–derived cardiomyocytes elucidate single-cell phenotype of brugada syndrome, J. Am. Coll. Cardiol. 68 (2016) 2086–2096.
[10] K. Penttinen, C. Prajapati, unpublished, 2021, .
[11] 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 (5) (2015), e0125366.
[12] R.P. Pölönen, K. Penttinen, H. Swan, et al., Antiarrhythmic effects of carvedilol and flecainide in cardiomyocytes derived from catecholaminergic polymorphic ventricular tachycardia patients, Stem Cell. Int. (2018) 2018.
[13] R.P. Pölönen, H. Swan, K. Aalto-Setälä, Mutation-specific differences in arrhythmias and drug responses in CPVT patients: simultaneous patch clamp and video imaging of iPSC derived cardiomyocytes, Mol. Biol. Rep. 47 (2020) 1067–1077.
[14] M. Juhola, K. Penttinen, H. Joutsijoki, K. Varpa, J. Saarikoski, J. Rasku, et al., Signal analysis and classification methods for calcium transient data of stem cell derived cardiomyocytes, Comput. Biol. Med. 61 (2015) 1–7.
[15] M. Juhola, H. Joutsijoki, K. Penttinen, K. Aalto-Setälä, Detection of genetic cardiac diseases by Ca2+ transient profiles using machine learning methods, Sci. Rep. 8 (2018) 9355.
[16] M. Juhola, H. Joutsijoki, K. Penttinen, K. Aalto-Setälä, Differentiation of genetic diseases on the basis of artificial intelligence, Eur. J. Biomed. Inform. 15 (3) (2019) 43–52.
[17] E.K. Lee, D.D. Tran, W. Keung, P. Chan, G. Wong, C.W. Chan, et al., Machine learning of human pluripotent stem cell-derived engineered cardiac tissue contractility for automated drug classification, Stem Cell Rep. 9 (2017) 1560–1572.
[18] C. Heylman, R. Datta, A. Sobrino, S. George, E. Gratton, Supervised machine learning for classification of the electrophysiological effects of chronotropic drugs on human induced pluripotent stem cell-derived cardiomyocytes, PLoS One 10 (2015), e0144572.
[19] D. Teles, Y. Kim, K. Ronaldson-Bouchard, G. Vunjak-Novakovic, Machine learning techniques to classify healthy and diseased cardiomyocytes by contractility profile, ACS Biomater. Sci. Eng. (2021), https://doi.org/10.1021/acsbiomaterials.1c00418.
[20] C. Coronnello, M.G. Francipane, Moving towards induced pluripotent stem cell-based therapies with artificial intelligence and machine learning, Stem Cell Rev. Rep. (2021), https://doi.org/10.1007/s12015-021-10302-y.
[21] H. Joutsijoki, K. Penttinen, M. Juhola, K. Aalto-Setälä, Separation of HCM and LQT cardiac diseases with machine learning of Ca2+ transient profiles, Method Inf. Med. 58 (2019) 167–178, 04/05.
[22] M. Juhola, H. Joutsijoki, K. Penttinen, D. Shah, K. Aalto-Setälä, On computational classification of genetic cardiac diseases applying iPSC cardiomyocytes, Comput. Methods Progr. Biomed. 210 (2021), 106367.
[23] O. Kramer, K-Nearest Neighbors, Dimensionality Reduction with Unsupervised Nearest Neighbors, Springer, Berlin, Heidelberg, 2013, pp. 13–23.
[24] A.J. Izenman, Linear Discriminant Analysis, Modern Multivariate Statistical Techniques, Springer, New York, NY, 2013, pp. 237–280.
[25] A. Tharwat, Linear vs. quadratic discriminant analysis classifier: a tutorial, Int. J. Appl. Pattern Recogn. 3 (2) (2016) 145–180.
[26] G. Bohling, Classical Normal-Based Discriminant Analysis, Technical Report, 2006. http://people.ku.edu/~gbohling/EECS833/Discrim.pdf.
[27] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, NY, 2009.
[28] C. Kwak, A. Clayton-Matthews, Multinomial logistic regression, Nurs. Res. 51 (6) (2002) 404–410.
[29] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. YuS, Z.H. Zhou, M. Steinbach, D.J. Hand, D. Steinberg, Top 10 algorithms in data mining, Known. Inf. Syst. 14 (2008) 1–37.
[30] L. Breiman, Random forests, Mach. Learn. 45 (2001) 3–32.
[31] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[32] S. Escalera, O. Pujol, P. Radeva, On the decoding process in ternary error-correcting output codes, IEEE Trans. Pattern Anal. Mach. Intell. 32 (1) (2010) 120–134.
[33] S. Escalera, O. Pujol, P. Radeva, Separability of ternary codes for sparse designs of error-correcting output codes, Pattern Recogn. Lett. 30 (3) (2009) 285–297.
[34] H. Joutsijoki, M. Haponen, J. Rasku, K. Aalto-Setälä, M. Juhola, Error-correcting output codes in classification of human induced pluripotent stem cell colony images, BioMed Res. Int. (2016), 3025057.
[35] V. Kecman, T.M. Huang, M. Vogt, Iterative single data algorithm for training kernel machines from huge data sets: theory and performance, in: L. Wang (Ed.), Support

Vector Machines: Theory and Applications, Springer, Berlin, Heidelberg, Germany, 2005, pp. 255–274.

[36] K.D. Garcia, C.B. de Sá, M. Poel, T. Carvalho, J. Mendes-Moreira, J.M.P. Cardoso, A.C.R.L.F. de Carvalho, J.N. Kok, An ensemble of autonomous auto-encoders for human activity recognition, Neurocomputing 439 (2021) 271–280.

[37] L. Sala, M. Bellin, C.L. Mummery, Integrating cardiomyocytes from human pluripotent stem cells in safety pharmacology: has the time come? Br. J. Pharmacol. 174 (21) (2016) 3749–3765. https://bpspubs.onlinelibrary.wiley.com/doi/full/10.1111/bph.13577.