Heart disease detection using an acceleration-deceleration curve-based neural network with consumer-grade smartwatch data

Naseri , Arman; Tax, David M.J.; Reinders, Marcel; van der Bilt, Ivo

DOI
10.1016/j.heliyon.2024.e39927

Publication date
2024

Document Version
Final published version

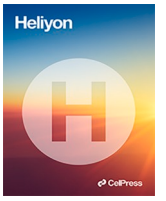
Published in
Heliyon

Heliyon 10 (2024) e39927

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Heart disease detection using an acceleration-deceleration curve-based neural network with consumer-grade smartwatch data

Arman Naseri [a,b,*], David M.J. Tax [b], Marcel Reinders [b], Ivo van der Bilt [a,c]

[a] Department of Cardiology, Haga Teaching Hospital, The Hague, Netherlands
[b] Pattern Recognition and Bioinformatics, Delft University of Technology, Delft, Netherlands
[c] Department of Cardiology, University Medical Center Utrecht, Utrecht, Netherlands

ARTICLE INFO

Keywords:
Mhealth
Ehealth
Wearables
Smartwatch
Machine learning
Atrial fibrillation
Heart failure

ABSTRACT

Cardiovascular disease (CVD) is the most important cause of morbidity and mortality worldwide. Early detection, prevention or even prediction is of pivotal importance to reduce the burden of cardiovascular disease and its associated costs. Low cost, consumer-grade smartwatches have the potential to revolutionize cardiovascular medicine by enabling continuous monitoring of heart rate and activity. When combined with machine learning(ML), the resulting large amounts of time series data hold the potential of detection, or exclusion of CVD. However, analyzing such large datasets is challenging due to the sparse presence of informative segments. Efficient selection of these segments is essential for developing predictive models for clinical deployment. The objective of this paper was to investigate the potential of an acceleration-deceleration curvebased ML model as a novel clinical indicator for the detection of cardiovascular diseases. We used data from the ME-TIME study; 42 participants from which 21 have a cardiovascular disease and 21 are health controls. Data from each subject was normalized to decrease inter-subject variability. A neural network model aggregated predictions per week. We showed that per-subject normalization by the peak value of curves during inactivity, aggregation of model predictions over a week, and using a contrastive loss, resulted in a predictive model with 99 % ± 3 % specificity and 40 % ± 49 % sensitivity on the development set, and 100 % specificity with 67 % ± 47 % sensitivity on the test set. Acceleration-deceleration curves are effective patterns for ruling out the presence of cardiovascular disease, but caution must be taken to properly preprocess the curves and carefully choosing a model that reduces the variability in the extracted curves.

1. Introduction

Cardiovascular disease (CVD), a leading cause of death [1], includes conditions like atrial fibrillation (AF) and heart failure (HF), which significantly contribute to the global morbidity- and mortality rate. To diagnose these conditions multiple tests like electrocardiograms (ECGs), ultrasound, cardiac CT and - MRI will be performed resulting in multiple hospital visits. These tests are often patient unfriendly and even potentially harmful. Additionally, significant costs are involved. In contrast, smartwatches have proven to

* Corresponding author. Department of Cardiology, Haga Teaching Hospital, The Hague, Netherlands.
E-mail address: a.naserijahfari@hagaziekenhuis.nl (A. Naseri).

https://doi.org/10.1016/j.heliyon.2024.e39927
Received 23 August 2024; Received in revised form 2 October 2024; Accepted 28 October 2024
Available online 30 October 2024

be easy to use, comfortable to wear and have the ability to continuously measure over long periods. The primary sensor in these devices is the photoplethysmogram (PPG), which is used to estimate heart rate [2]. Heart rate variability (HRV) [3] is a strong clinical indicator of heart disease, positioning smartwatches as potential medical informative devices for consumers. This potential is further enhanced by the ongoing integration of additional sensors into smartwatches, like accelerometers, which can provide more context to heart rate (variability) data, thereby improving its predictive power.

Machine learning (ML) models are crucial to detect and predict cardiovascular outcomes from the resulting (HRV) data. These models not only aid in early detection but can also be useful in monitoring the possible progression of heart-related conditions over time and/or treatment effects. The strength of ML lies in its ability to automate the analysis of the vast amounts of HRV data. Furthermore, the data is not clinically informative to a physician at first sight, yet HRV patterns related to CVD can be effectively discovered and utilized by a ML model.

Recent success in detecting atrial fibrillation (AF) using smartwatch data from raw PPG and ECG signals demonstrates the potential of wearable devices. However, these studies have primarily relied on short-term recordings [4–6], whereas the true potential of smartwatches lies in continuous, long-term outpatient monitoring and analysis. Similarly, heart failure detection has shown promising results [7], but these findings are also based on short ECG recordings during hospitalization.

Moreover, the raw data and algorithms used in large studies led by manufacturers [8–11] are proprietary and device-specific, making it difficult to benchmark these models in subsequent studies and impeding cross-device algorithm development for a wider range of cardiovascular diseases (CVDs).

In addition, some studies have utilized PPG-derived heart rate data [12,13], sometimes in combination with step counter data [14]. However, these models require ECG-based labels for training, limiting their flexibility to be adapted for other CVDs.

In contrast, our method demonstrates a robust cardiovascular health indicator for both atrial fibrillation and heart failure. It only requires heart rate data—whether from PPG or ECG—and step counter data, which is readily available from commercial smartwatches. Importantly, our approach does not depend on labels from other wearables that need to be worn or implanted simultaneously, offering greater flexibility for use in different contexts.

Entrusting the entire longitudinal data set to a machine learning model can be counterproductive if only a small proportion of the time series contains indicators for cardiovascular disease.

Therefore, it is beneficial to select a subset of informative segments from the time series, guided by pre-existing cardiological knowledge. This can benefit the model's performance and provide more insights to physicians by relating the outcome of the model with patterns known from the field.

In exercise physiology an important prognosticator is the maximum aerobic capacity (VO2max) or maximum oxygen consumption. It is associated with lower cardiovascular and total mortality. Heart rate recovery (HRR) after exercise [15–17] is associated with VO2max and easily measurable using wearables. It refers to the speed at which the heart rate returns to a baseline level after a physical activity. A rapid HRR is typically indicative of a healthier heart and a more efficient autonomic nervous system. Conversely, a delayed HRR is associated with a higher risk of cardiovascular morbidity and mortality.

This type of controlled testing is typically not present in observational, longitudinal data as subjects are monitored in their daily free-living environment. Nonetheless, similar patterns to the HRR after exercise can occur in daily life, for example due to physical activity like exercising, climbing stairs or household chores, certain foods or beverages [18,19] and emotional responses. These occurrences essentially serve as a rudimentary gauge of cardiovascular fitness, akin to a 'poor person's exercise test', reflecting the body's response to routine stressors and activities.

In addition to the HRR after exercise, heart rate acceleration capacities is associated with heart failure [20].

This paper explores patterns characterized by both an onset phase, where the heart rate increases to a peak level, and a recovery phase, where the heart rate recovers back to a baseline level. We introduce the concept of 'acceleration-deceleration (acc-dec) curves' to describe these patterns. Our contributions include a detector to extract acc-dec curves from heart rate time series data. Furthermore,

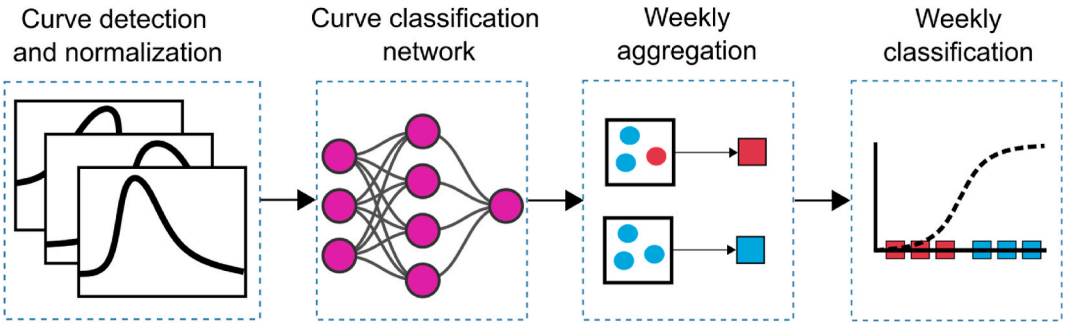


Fig. 1. Proposed model to detect cardiovascular disease from acc-dec curves. In the first stage, the curve detector extracts acc-dec curves and applies normalization per subject. These are then grouped in batches (both during training and deployment) and input to a neural network, which assigns a probabilistic score ranging from zero to one to each curve, illustrated as blue (REF) and red (CVD) circles. Subsequently, these predictions are grouped per subject by week, illustrated by black squares. An average of these weekly grouped predictions is then calculated, represented by blue and red squares. In the final stage a classifier is trained on the weekly aggregated predictions.

we present a ML model capable of excluding cardiovascular disease (CVD) based on acc-dec curves from subjects with CVD and reference (REF) subjects without CVD, utilizing a normalization based on a step counter, weekly aggregation of curves, and a divergence-based loss for improved accuracy. We also demonstrate the clinical relevance of our model through validation on the ME-TIME [21] data set, showing its potential in enhancing early detection and monitoring of cardiovascular conditions.

2. Materials and methods

The complete ML model pipeline is illustrated in Fig. 1 and detailed in further sections. Briefly, acc-dec curves are extracted from the heart rate time series and normalized individually for each subject. The data is then grouped in batches where each batch contains curves from different patients and multiple weeks. The batching is required, because the outputs of the neural network are aggregated on a weekly basis in a later phase. Therefore, the number of curves must span an entire week to perform meaningful aggregation. A neural network then assigns a score between 0 and 1 to each curve in a batch, where larger than 0.5 is classified as coming from a patient with cardiovascular disease (CVD), and smaller than 0.5 is classified as healthy patient (REF). After all curves in a batch have been assigned a prediction by the network, the scores of predictions that fall in the same week, are averaged per subject together and finally a logistic layer classifies the weeks as belonging to either the REF or CVD class.

2.1. Curve detection and normalization

The data is first filtered by an eighth order Butterworth low-pass filter with critical frequency of 0.075 Hz and this filter is applied both forward and backward to cancel out phase delays/shifts introduced by the filter [22]. The critical frequency was chosen based on visual inspection compared to the raw data, ensuring that the filter preserved the overall shape of the curve while effectively attenuating higher frequency noise.

Then, the detector extracts acc-dec curves from the heart rate time series by identifying three fiducial points as demonstrated in Fig. 2. First, the peak heart rate of the acc-dec curve is identified. This peak heart rate is the absolute heart rate value at the highest point of the acc-dec curve, with the corresponding peak point indicating its specific location. It is identified using SciPy's local maxima detection method [23], with a key parameter of this method being the prominence parameter. Prominence measures the relative height of the peak heart rate compared to the minimum heart rate in the surrounding data. It is calculated by finding the difference between the peak heart rate and the lowest heart rate value in the region between this peak and the adjacent peaks on either side.

Multiple acc-dec curves are identified for each subject, aligned at their peak points, and then normalized for amplitude. Two normalisations are considered: a quantile normalization, and a mean inactive-peak normalization. In the quantile normalization, for each subject, the peak values that fall within the q-th quantile are averaged and used as normalization constant. In the mean inactive-peak normalization, for each subject, all peaks are averaged for which the subject is inactive (i.e. step counter is 0) during the time interval between the onset and recovery point of the acc-dec curve, and that is used as normalization constant.

2.2. Curve classifier

To maintain class-balance for the neural network model, during training the curves are randomly resampled, such that there is an equal number of curves originating from both REF and CVD subjects. In order to have samples of the same dimensionality, we consider acc-dec curves from 300 s before and until 600 s after the peak. Since the heart rate sample frequency is 0.2 Hz, this results in samples of dimensionality 241 ((300 + 600 + 5) *0.2), where the additional 5 s represents the peak itself. The acc-dec curves are then grouped in batches of 8192 curves, totalling 6 batches and used as input to a neural network.

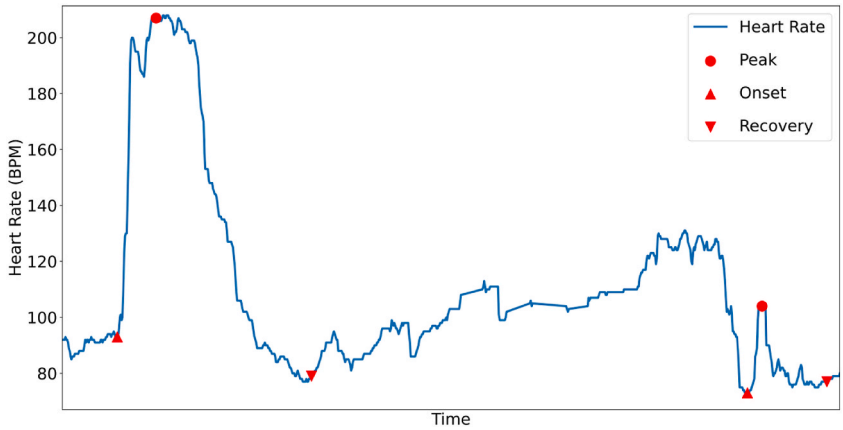


Fig. 2. Example of a heart rate signal in which two acc-dec curves are detected. Red circle, upward arrow and downward arrow represent peak, onset and recovery fiducial points, respectively.

The neural network model consists of a (acc-dec) curve classifier, which classifies each acc-dec curve as belonging to the CVD class or REF class, followed by per-subject weekly aggregation and weekly classification. During training, backpropagation starts at the weekly classification stage and flows backward through the network, reaching all the way to the curve classifier.

The curve classifier consists of two layers, where the first layer maps the 241 samples to a 20-dimensional hidden space with a tanh activation function, whereafter the final layer maps the 20-dimensional representation to a single output with a logistic activation function. This output, which varies from 0 to 1, represents a confidence score assigned by the model. A score of 1 indicates the model's highest confidence in classifying the curve as belonging to the CVD class, while a score of 0 corresponds to its association with the REF class.

2.3. Weekly aggregating neural network

For each subject, the outputs of the curve classifier from the same week are averaged. This average essentially serves as an indicator of the prevalence, or lack thereof, of CVD within that specific week for the individual subject. This results in a distribution representing weekly aggregated scores over all subjects. It should be emphasized that in order to enable weekly predictions, predictions based on weekly-aggregated acc-dec curves must be performed also during deployment of the model. Furthermore, it is assumed that during training all peaks from a CVD subject are considered to be CVD and all peaks from a REF subject are considered non-CVD. The same holds for the weekly aggregated scores.

Finally, a logistic regressor is trained on the weekly averaged data, again outputting a score between 0 and 1, for which, again, values between 0 and 0.5, are predicted to be non-CVD and between 0.5 and 1, predicted as CVD.

2.4. Divergence-based loss

The model is optimized by minimizing the cross-entropy between the weekly aggregated predictions $\hat{y}$ and the label y of the subject it originated from.

Additionally, we used different regularization functions based on divergence methods that maximize the dissimilarity between predicted REF and CVD outcomes. This results in a loss function L:

$$L = L_{CE}(y, \hat{y}) + \alpha D(\hat{y}_{CVD}, \hat{y}_{REF}) \tag{1}$$

where L_{CE} is the cross-entropy loss, D the divergence-based regularizer and α the regularization coefficient controlling the impact of the regularizer on the overall loss function.

The first regularization method aims to only separate the means and is formulated as:

$$D_{mean}(\hat{y}_{CVD}, \hat{y}_{REF}) = \frac{1}{(m(\hat{y}_{CVD}) - m(\hat{y}_{REF}))^2} \tag{2}$$

where $m(\hat{y})$ is the sample mean of the weekly predictions of the corresponding class.

In the second method, we separate the variances in addition to the means using a KL-divergence assuming Gaussian distributions [24]. In the univariate case this becomes:

$$L_{KL}(\hat{y}_{CVD}, \hat{y}_{REF}) = log\left(\frac{s(\hat{y}_{REF})}{s(\hat{y}_{CVD})}\right) + (s(\hat{y}_{REF}))^2 + \frac{(m(\hat{y}_{CVD}) - m(\hat{y}_{REF}))^2}{2(s(\hat{y}_{REF}))^2} \tag{3}$$

where $s(\hat{y})$ is the sample standard deviation of the weekly predictions of the corresponding class.

In contrast to the KL-divergence, we aim to maximize the divergence. Therefore in analogy with the mean divergence in eq. (2) we take the reciprocal:

$$D_{KL} = \frac{1}{L_{KL}} \tag{4}$$

In the final method, we consider the contrastive loss [25]. This loss considers the Euclidean distance between all possible pairwise combinations of the weekly predictions, $\hat{y}$. Pairs belonging to the same class are identified as positive pairs, whereas those from different classes are termed negative pairs. For the positive pairs, the Euclidean distance is defined as:

$$d_k^{pos} = ||\hat{y}_i - \hat{y}_j||_2^2 \tag{5}$$

where d_k^{pos} contains the Euclidean distances of all positive pairs and similarly d_l^{neg}, denotes those of all negative pairs.

The objective is to keep positive pairs closely together, while simultaneously separate negative pairs, thereby achieving a clear distinction between the classes. The contrastive loss is given as:

$$D_{contrastive} = \frac{1}{N^{pos}} \sum_{i=1}^{N^{pos}} \left(d_k^{pos}\right)^2 + \frac{1}{N^{neg}} \sum_{l=1}^{N^{neg}} max\left(0, \beta - \left(d_l^{neg}\right)^2\right) \tag{6}$$

The first term of the equation aims to reduce the Euclidean distances between positive pairs, denoted by N^{pos} (i.e. the total number

of these pairs). Conversely, the second term seeks to increase the distance between negative pairs, where N^{neg} represents the total number of negative pairs. Additionally, the introduction of a margin hyperparameter, β, combined with the use of the max operator, plays a crucial role. It limits the penalty to distances up to β, allowing the model to concentrate on "hard examples" (i.e. negative and positive pairs that are much alike). This focus is beneficial as it prevents unnecessary optimization of samples that are already satisfactorily positioned, where further re-positioning wouldn't enhance the model's performance.

2.5. ME-TIME data set

The ME-TIME study (registered at clinicaltrials.gov with ID NCT05802563) is an observational, longitudinal study conducted at the Haga teaching hospital and is approved by the Institutional Review Board (or Ethics Committee) of METC-LDD (protocol code NL73708.058.20) for studies involving humans. The study, which included a 12-week observation period, aims to develop methods and algorithms for remote detection and prevention of cardiovascular disease using Fitbit smartwatches [21]. Tables 1 and 2 show the characteristics of the development set, used to train and tune a neural network model, as well as the test set.

The REF group consists of 15 subjects in the train and 6 subjects in the test set. The CVD group is divided into train and test in a similar way.

In comparison to the REF group, the CVD group consists of slightly older individuals with higher BMI and a higher prevalence of hypertension. The CVD group consists of subjects with some type of atrial fibrillation (AF) or heart failure (HF).

2.6. Evaluation

The measures reported are the Sensitivity (a.k.a. True Positive Rate) and Specificity (a.k.a. True Negative Rate), defined as

$$Sensitivity = \frac{TP}{TP + FN} \tag{7}$$

$$Specificity = \frac{TN}{TN + FP} \tag{8}$$

where True Positive (TP) represents the number of weeks correctly classified as CVD. False Negative (FN) the number of weeks misclassified as REF. False Positive (FP) the number of weeks misclassified as CVD. True Negative (TN) the number of weeks correctly classified as REF.

This results in a sensitivity and specificity for each subject after which the population average and standard deviation across subjects is reported.

2.7. Model training

Parameters such as the margin and regularization coefficient mentioned in Section 2.4 cannot be determined through the neural network's learning process and are called hyperparameters. Instead, these are found through trial-and-error by trying out different values. The best hyperparameter values are determined by selecting the configuration with the highest average accuracy across folds in a leave-6-subjects-out stratified cross-validation approach on the training set only. This procedure is further elaborated in Appendix A.

Table 1
Characteristics of the development set used to train and tune a neural network model.

Characteristic		REF	HF	PersAF	PermAF	HF + AF
Participants, n		15	8	3	4	15
Age, y	18–39	6	0	0	0	6
	40–54	5	1	1	0	2
	55–64	4	3	0	0	3
	>65	0	4	2	4	10
Sex	Male	4	6	3	2	11
	Female	11	2	0	2	4
BMI	18.5–24.9	7	2	0	1	3
	25–29.9	5	2	2	2	6
	>30	3	4	1	1	6
Diabetes	Yes	0	4	0	2	6
	No	15	4	3	2	10
Smoking	Yes	0	5	0	3	8
	No	15	3	3	1	7
Hypertension	Yes	1	5	1	3	9
	No	14	3	2	1	6
Device	Charge 5	10	5	2	3	10
	Inspire 2	5	3	1	1	5

Table 2
Characteristics of the test set used to evaluate the performance of the trained and tuned model.

Characteristic		REF	HF	PersAF	PermAF	HF + AF
Participants, n		6	4	1	1	6
Age, y	18–39	4	0	0	0	0
	40–54	1	1	0	0	1
	55–64	1	1	0	0	1
	>65	0	2	1	1	4
Sex	Male	3	3	1	0	4
	Female	3	1	0	1	2
BMI	18.5–24.9	3	2	0	0	2
	25–29.9	3	1	1	1	3
	>30	0	1	0	0	1
Diabetes	Yes	0	1	1	0	2
	No	6	3	0	1	4
Smoking	Yes	0	1	0	0	1
	No	6	3	1	1	5
Hypertension	Yes	0	2	1	0	3
	No	6	2	0	1	3
Device	Charge 5	5	4	0	1	5
	Inspire 2	1	0	1	0	1

3. Results

3.1. Inter-subject and intra-subject variability

The results illustrated in Fig. 3 highlight significant variations in the distribution of peak heart rates within the acc-dec curves among participants. For this analysis, we also include a paroxysmal AF group. These variations are evident not only between different subjects (inter-subject variability) but also within the measurements from individual subjects (intra-subject variability). By applying the mean inactive peak normalization and sorting the subjects by median peak heart rate, it becomes apparent in Fig. 3 that the normalized peak heart rate serves as an informative feature for CVDs. Specifically, values approximately above 1.1 (10 percent above the baseline peak heart rate) generally correspond to reference subjects. Furthermore, values below the 1.0 baseline seem to correspond to heart failure patients. These findings underscore the complex nature of cardiovascular responses as captured by acc-dec curves.

We investigated the impact of the normalization on the difference in acc-dec curves between the CVD and REF group. Fig. 4 shows average acc-dec curves, with and without normalization. Here, the distance between the REF and CVD curves appears to be larger when using the normalization by the mean inactive-peak. As we strive towards differentiating between these groups, we selected for this normalization scheme. Appendix B demonstrates that other quantile values perform worse. We also investigated the impact of the prominence and activity level in Appendix C.

This revealed that, using acc-dec curves with mean inactive-peak normalization, a prominence of 20 BPM and activity levels above 20 steps, the distance between the average REF and CVD curve appears to be largest, among the methods that we investigated. Therefore we use this configuration to train a ML model.

3.2. Ablation study of predictive model

Using these hyperparameters, we developed ML models to classify acc-dec curves of participants into CVD patients or REF individuals.

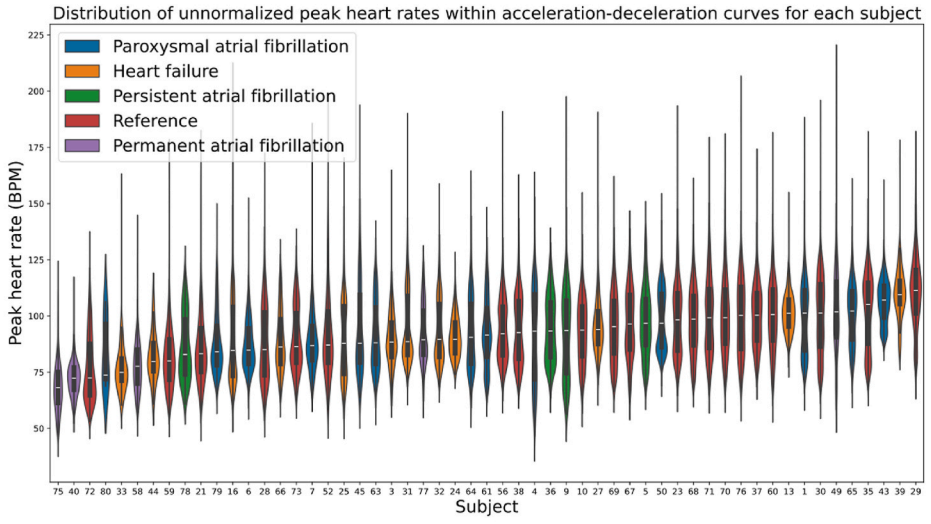
We performed an ablation study (shown in Table 3) starting with a naive baseline, which only consists of the curve-based classifier without normalization or weekly aggregation/classification from Fig. 1. Then we have sequentially added the mean inactive-peak normalization and aggregation, and the last component added is the divergence-based regularization, for which we investigate three alternatives: mean divergence, KL divergence and contrastive loss. By doing so, the impact of each of these additions on the model's performance is quantified.

The results in Table 3 show that the naive model performs poorly. Applying mean inactive-peak improves the specificity of the model. This enhancement becomes more pronounced when the aggregation technique are also implemented.

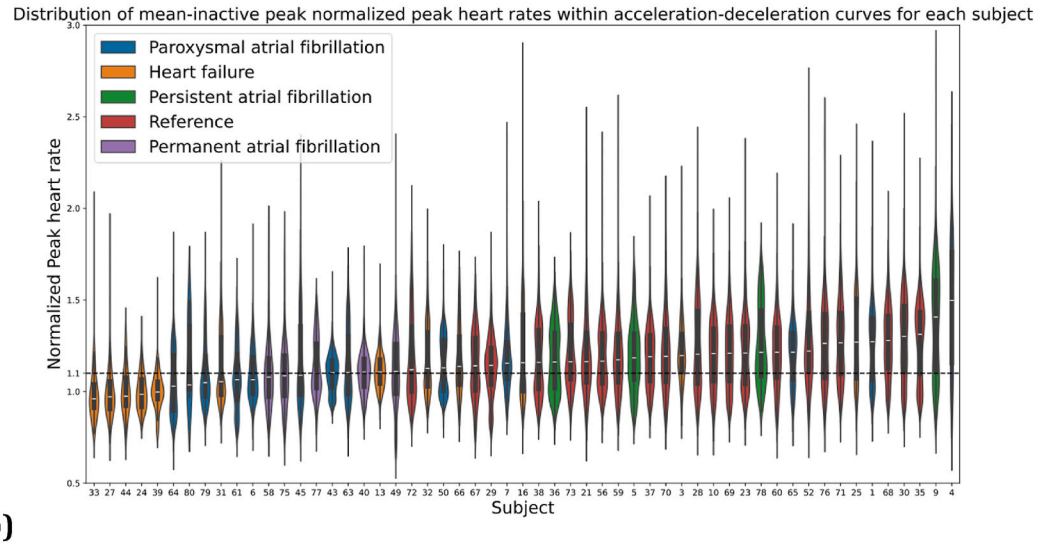
On top of normalization and aggregation, the addition of the divergence losses is considered, resulting in three alternatives.

When mean divergence is applied, the model's performance significantly deteriorates. Closer examination of the mean divergence model's predictions in Fig. 5b reveals that it always classifies all samples as CVD, resulting in 100 % sensitivity at 0 % specificity.

The specificity of the model benefits most from contrastive regularization. While each added component increases specificity, it simultaneously leads to a reduction in sensitivity. Considering the high specificity, the model is very good at detecting reference cases when they are present. However, the low average for the sensitivity (43 %), and its high standard deviation (48 %), indicate that the model is not consistently good at identifying CVD cases as seen in Fig. 5b. Specifically, for some CVD test subjects, it classifies almost all positive weeks correctly, while in others, it fails to classify any weeks correctly. Thus, the model's positive predictive performance closely resembles that of random guessing, underscoring its lack of consistency and reliability in detecting CVD cases. In addition,



(a)



(b)

Fig. 3. Unnormalized (top, a) and mean-inactive peak normalized (bottom, b) distribution of peak heart rates per subject. A: Distribution of unnormalized peak heart rate in acc-dec curves for each subject. B:Distribution of mean-inactive peak normalized peak heart rate across acceleration-deceleration curves for each subject.

contrastive regularization results in an accuracy of 85 %, positive predictive value of 100 % and negative predictive value of 82 %.

Finally, KL divergence performs better than mean divergence, however it does not improve upon using contrastive regularization or even only normalization and aggregation. However, Fig. 5a reveals that it performs best in terms of the area under the receiver operator curve.

4. Discussion & conclusion

We have introduced acc-dec curves in combination with a ML model to demonstrate high specificity, reliably identifying the absence of cardiovascular disease in longitudinal data acquired from consumer-grade wearables. Deployed as a "rule out" modality this can be very powerful in clinical practise. Currently, many patients visiting a physician with complaints require further testing to diagnose the disease causing these complaints. Medical tests are often designed to have a high specificity to optimize the negative predictive value for the disease. For example, cardiac CT has a 99 percent specificity for the presence of coronary calcification [26] making it a powerful tool to rule out coronary artery disease and strongly modifying cardiovascular risk. However, performing an expensive scan (which also uses harmful ionizing radiation and contrast causing patients' discomfort) only provides information on the

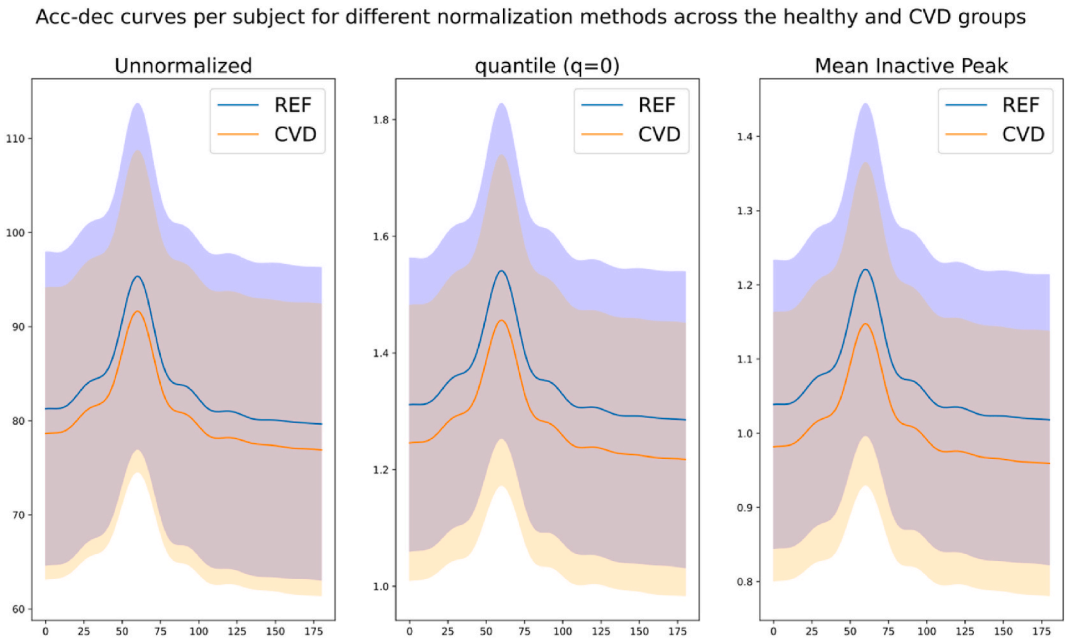


Fig. 4. Unnormalized (left), quantile normalization (middle, using q = 0, effectively using the smallest peak) and mean inactive-peak normalization (right).

Table 3
Ablation analysis of average model performance across subjects with incremental addition of: mean inactive-peak (M) normalization, Weekly Aggregation (A), Mean Divergence (MD), KL Divergence (KL), Contrastive Loss (CL). Naive model consists of curve-based classifier only. Standard deviations are provided in parentheses to illustrate the variability across subjects.

Set	Metric	Naive	M	M + A	M + A + MD	M + A + KL	M + A + CL
Development	Sensitivity	44 (28)	60 (25)	44 (48)	20 (40)	51 (48)	40 (49)
	Specificity	63 (18)	61 (19)	88 (16)	80 (40)	89 (27)	99 (3)
Test	Sensitivity	61 (18)	64 (19)	67 (48)	0 (0)	67 (47)	67 (47)
	Specificity	53 (19)	62 (12)	83 (19)	100 (0)	79 (37)	100 (0)

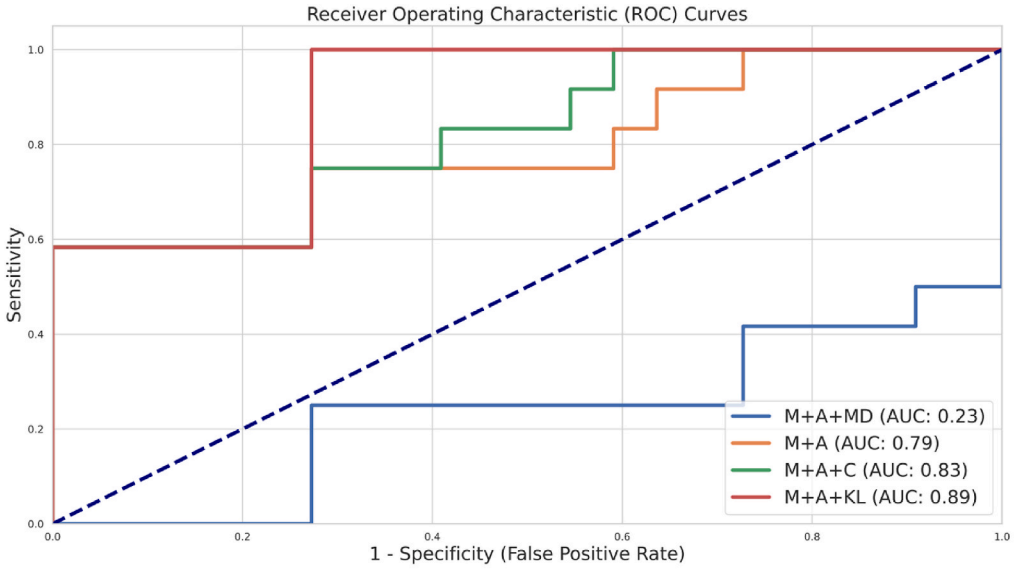
current status (a 'snap shot'). In contrast, a monitoring system based on this ML model can be used by the patients for an extensive period of time, has no harmful effects on the patient and it also has the potential to avoid unnecessary visits to a cardiologist for patients who have symptoms that are initially suspected to be related to heart problems, but after assessment are determined not to be. Our method on predicting CVD also has a high specificity, for which a similar clinical efficiency reasoning can be followed.

Secondly it is shown that it is important that the acc-dec curves are normalized per-subject based on scaling by the step count. In our analysis, scaling the curves using the average peak value observed during inactivity proves to be effective. Our study involved comparing acc-dec curves between healthy individuals and CVD patients, focusing on three distinct ranges based on step counter readings: no steps (0), low activity (1–20 steps), and higher activity (exceeding 20 steps). Curves recorded, during periods of higher activity appeared to be more effective in differentiating between the two groups.

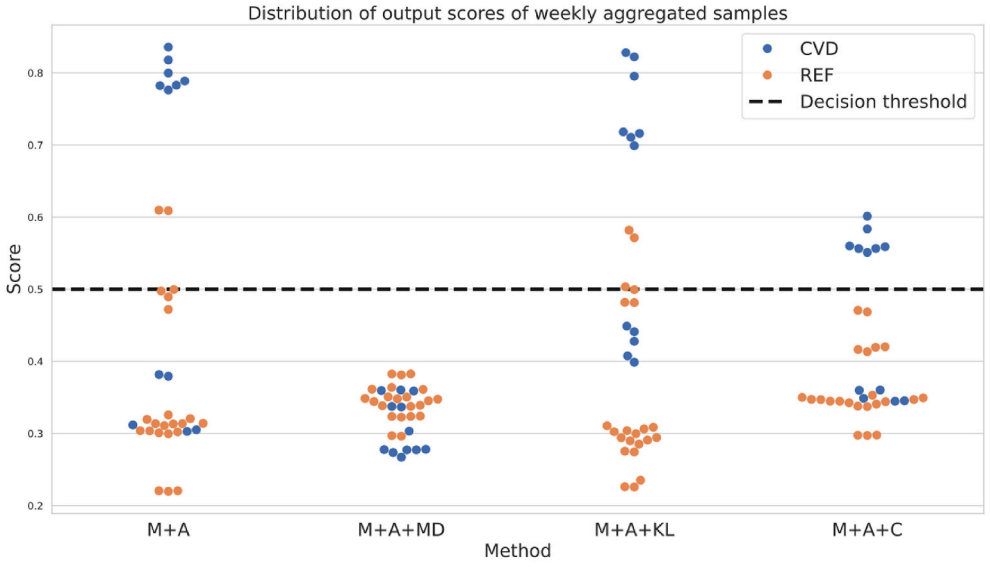
Thirdly, we show that it is beneficial to aggregate model predictions of curves that fall within the same week. The effectiveness may be attributed to its ability to mitigate the large inter- and intra-subject variability in acc-dec curves. Such variabilities could challenge machine learning models to generalize effectively across a diverse population and must generally be accounted for. Other time intervals could also be explored and in general, incorporating information about time (of the day) could be beneficial.

Fourthly, we show that it is effective to use a divergence-based loss that ensures that the model finds a representation where data from the two classes, healthy and cardiovascular disease, are far apart. Since originally the mean divergence and KL-divergence are suited for minimization problems, we took the reciprocal to maximize the distance between the two classes. However, other functions that transform divergence-based losses into maximization methods could also be explored such as negation. Compared to the reciprocal functions, the contrastive loss performs best. This difference may be because the mean and KL divergence only optimize for summary statistics (mean and variance), whereas contrastive loss penalizes pairwise predictions, thus operating at a more granular level. Other contrastive functions such as the triplet loss could also be explored.

The results in Table 3 and Fig. 5b show that the models are useful to rule out the presence of CVD, due to the high specificity (100 %), but cannot detect CVD very well due to the low sensitivity (65 % ± 47 %). Because the models have this characteristic, they can be a very strong tool in medicine, and especially in cardiology, as clinicians usually work "per exclusionem" or by ruling out a diagnosis.



(a)



(b)

Fig. 5. ROC curves for aggregation methods (top, a) and output scores of weekly aggregated predictions on the test set (bottom, b). Colors represent the true labels: blue for CVD and orange for REF. The dashed line indicates the decision threshold of 0.5, with scores below 0.5 classified as REF and scores above 0.5 classified as CVD.

Therefore may be of significant clinical importance.

Our study is limited by the fact that the sample size on which the model is tested is rather small.

Lastly, we have considered the classification performance of the weekly aggregated data. By utilizing subject labels to give predictions on a per-subject basis (i.e. aggregating over multiple weeks), this approach might allow for more consistently accurate predictions.

Future research should be directed towards external validation of the model on various datasets (e.g. multi-center study) from different smartwatch devices to confirm the data. The goal should be to safely implement these models in clinical practice and to

improve medical care. Furthermore, more diagnoses should be trained using the methods we propose and this work should be directed towards diseases with a high prevalence and incidence such as coronary artery disease.

Furthermore, successful integration of such models into clinical practice comes with challenges. For example, seamless integration with existing healthcare systems is essential for the model to provide real-time, actionable insights. Furthermore, healthcare providers would need training on interpreting model outputs and effectively integrating them into patient care workflows. Patient compliance also plays a critical role, in order to capture sufficient acc-dec curves. Additionally, a proper database infrastructure is required to store and manage the data effectively, and cloud platforms are a suitable option for this purpose. However, the use of cloud platforms introduces privacy considerations that must be addressed to ensure the protection of sensitive patient information. Addressing these challenges is vital to ensure smooth adoption and widespread implementation of the proposed model.

In summary, this study demonstrates the effectiveness of acc-dec curves acquired from a consumer-grade smartwatch combined with machine learning, offering a non-invasive, efficient, and powerful tool to rule out cardiovascular disease. This shows promise of transforming clinical practice and enhancing patient care through advanced, ML-driven methodologies, contributing to a new era of remote patient monitoring.

CRediT authorship contribution statement

Arman Naseri: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **David M.J. Tax:** Writing – review & editing, Supervision. **Marcel Reinders:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Ivo van der Bilt:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Conceptualization.

Informed consent statement

Informed consent was obtained from all subjects involved in the study.

Data sharing

The anonymised heart rate, step counter and deidentified timestamps of the Fitbit timeseries data are available on request to a.naserijahfari@hagaziekenhuis.nl. Data will be shared with publication including descriptions of each field and instructions, after approval of a project proposal, with a signed data access agreement. The code will be made available on https://github.com/Armannas/acc-dec-curves.

Institutional review board statement

The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of METC-LDD (protocol code NL73708.058.20) for studies involving humans.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the first author used chatGPT in order to correct spelling, grammar and to improve sentence clarity. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.}

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Validation and hyperparameter tuning

Utilizing gridsearch and Stratified Leave-6-Subjects-Out Cross-Validation (SL6SOCV) on the development set, we fine-tuned the model's hyperparameters [27]. Stratification maintains an equal number of Ref and CVD subjects in each training and validation fold to prevent sampling bias during training and evaluation of the model. The leave-6-subjects-out component guarantees that all samples originating from the same subjects are grouped within a single fold. This prevents information leakage about the subjects into the validation set.

The hyperparameter space is shown in Table A1 and all possible combinations were explored. The combination with the highest average accuracy using SL6SOCV is used to train a model on the development set, whereafter this model is evaluated on the test set.

Table A1
Gridsearch hyperparameters explored and the corresponding optimal values for each method. mean inactive-peak (M) normalization, Weekly Aggregation (A), Mean Divergence (MD), KL Divergence (KL), Contrastive Loss (CL). Naive model consists of curve-based classifier only.

Hyperparameter	Values	Description	Naive	M	M + A	M + A + MD	M + A + KL	M + A + C
w_on	300	Search window onset	300	300	300	300	300	300
w_recovery	[600, 900]	Search window recovery	600	600	600	600	600	600
n_hidden	[100, 20, 10]	Neurons in last hidden layer	10	10	10	10	10	10
n_fc	[1,2,3]	Fully-connected layers	2	2	2	2	2	2
n_conv	[0,1,2]	Convolutional layers	0	0	0	0	0	0
act_funcs	[Tanh, RELU]	Activation function	Tanh	Tanh	Tanh	Tanh	Tanh	Tanh
b_norm	[True, False]	Batch norm	0	0	0	0	0	0
Dropout	[0, 0.6]	Dropout	0	0	0	0	0	0
Lr	[0.02, 0.05]	Learning rate	0.05	0.02	0.05	0.05	0.05	0.05
batch_size	8192	Batch size	8192	8192	8192	8192	8192	8192
kernel_size (conv only)	[5, 20]	Kernel size	N/A	N/A	N/A	N/A	N/A	N/A
pool_size (conv only)	2	Pooling size, stride 2	N/A	N/A	N/A	N/A	N/A	N/A
n_channels (conv only)	16	Conv channels	N/A	N/A	N/A	N/A	N/A	N/A
w_agg_size	604800	Aggregation window (seconds)	604800	604800	604800	604800	604800	604800
w_agg_stride	259200	Aggregation window stride (seconds)	259200	259200	259200	259200	259200	259200
Alpha	[1e-5,1e-4,1e-3,1e-2,1e-1,1,10]	Divergence loss regularization coefficient	N/A	N/A	N/A	1e-5	1e-5	1e-2
Beta	[0, 0.1, 1, 10]	Contrastive loss margin coefficient (contrastive only)	N/A	N/A	N/A	N/A	N/A	0.1

The number of neurons in each fully-connected hidden layer linearly decreases based on the input and last hidden layer n_{hidden} according to the formula:

$$N_l^n = n_{hidden} + \left\lceil \frac{\frac{w_{on}+w_{recovery}+5}{5} - n_{hidden}}{n_{fc}+1} \right\rceil \cdot l, l = 0, \ldots, n_{fc}+1 \tag{A1}$$

where N_l^n is the number of neurons in the l-th hidden layer. For example, an input size of 160 neurons, with 10 neurons in the last hidden layer and 2 hidden layers in between would result in a network with 110, 60 and 10 neurons in the first, second and last hidden layer respectively.

The fully-connected layers are considered with and without preceding convolutional layers. Each convolutional layer's output is:

$$N_l^k = N_{l-1}^f - w^k + 1 \tag{A2}$$

where N_l^k is the feature map size after applying the kernel, N_{l-1}^f is the feature map size (or number of neurons) of the previous layer and wˆ k is the kernel size.

Similarly, pooling is applied after applying the kernel with a stride of 2:

$$N_l^f = \frac{N_l^k - w^p}{2} + 1 \tag{A3}$$

Appendix B. Sensitivity analysis on quantile normalization

Figure B1 illustrates quantile normalizations for different quantile values. The minimum (q = 0) is perhaps similar within a group but different between the REF and CVD groups. This would make sense as resting heart rate of patients with CVD is characteristically different from healthy subjects. In contrast, higher quantiles that ignore extremeties, do not capture this and are more similar across groups.

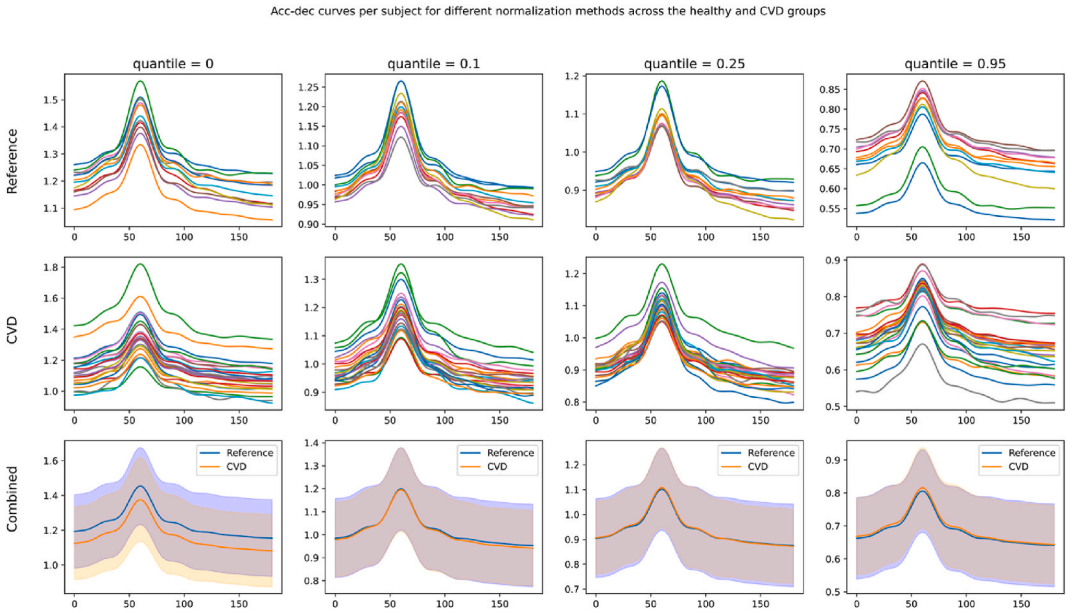
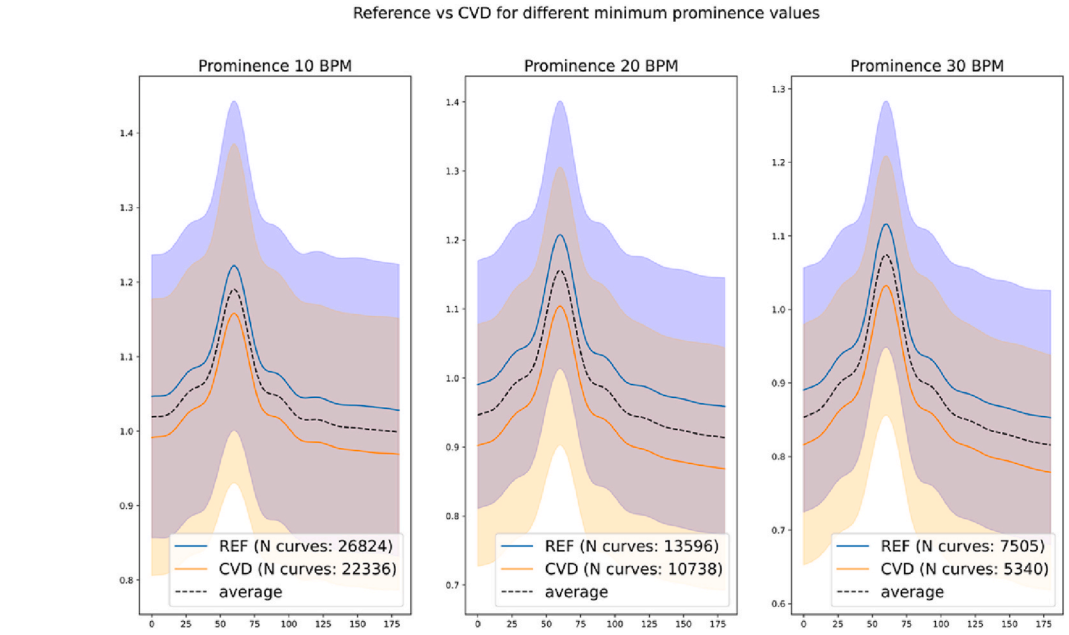


Fig. B1. Effect of quantile normalization applied per-subject. Every curve in the first two rows represent the average curve from each subject of the Ref and CVD group, respectively. The last row show the average curve per group.

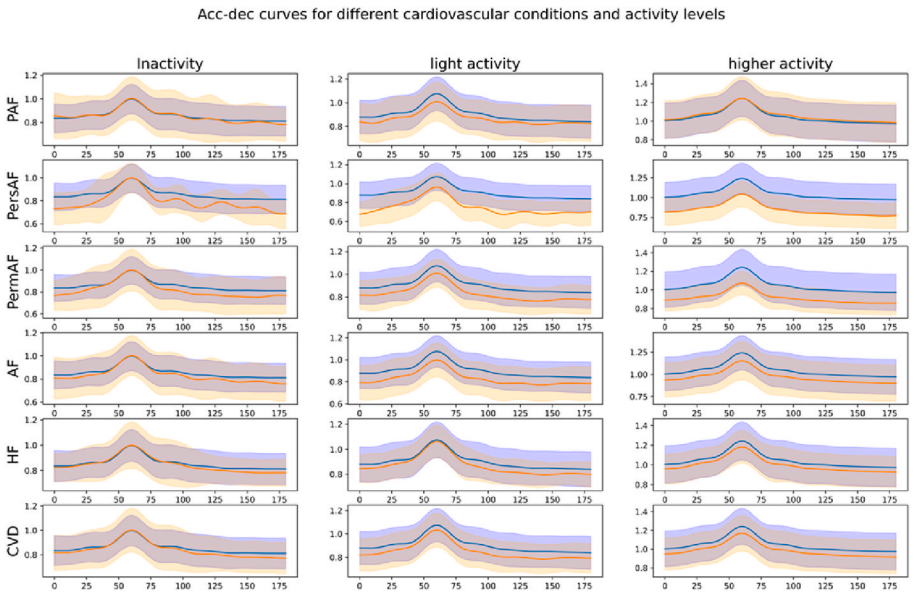
Appendix C. Impact of prominence and activity levels

The acc-dec curves are derived from the peak detection algoritm which uses the prominence value to determine whether a peak stands out. To investigate the influence of how much the peak stands out, Figure \ref{fig:prominences} shows the average curve for the REF and CVD class for three different different values of the prominence. A prominence level that is either too small (10 BPM) or too high (30 BPM) appears to reduce the distance between the average REF and CVD curve, especially noticeable when comparing either group to the overall mean curve in the region before and after the peak, while 20 BPM is a reasonable compromise.

Next we analysed the influence of activity level on the difference between the CVD and REF group. Figure C1 shows that, compared to peaks during inactivity (0 steps) and light activity(1–20 steps), difference between acc-dec curves of the CVD and REF group is largest when considering only higher activity (>20 steps). This is true for all CVDs, except for Paroxysmal Atrial Fibrillation (PAF) indicating that it is a more suitable configuration for a ML model.



(a)



(b)

Fig. C1. Effect of a prominence (top, a) and activity level (lower, b) on the acc-dec curves for the REF (blue) and CVD (orange) class. The curves are averaged per class with their corresponding + - 2 standard deviations interval. A:Average REF and CVD curves with a prominence value of 10 (left), 20 (middle) and 30 (right) BPM. Black dashed line denotes average over all curves. B: Effect of different levels of activity. Inactivity, light activity and higher activity correspond to 0, 0 to 20 and higher than 20 recorded steps, respectively. CVDs investigated are PAF (Paroxysmal Atrial Fibrillation), PersAF (Persistent Atrial Fibrillation), PermAF (Permanent Atrial Fibrillation) and HF (Heart Failure). AF (Atrial Fibrillation) contains PAF, PersAF and PermAF. CVD contains all CVDs.

References

[1] A. Shi, Z. Tao, P. Wei, J. Zhao, Epidemiological aspects of heart diseases, Exp. Ther. Med. 12 (2016) 1645–1650.
[2] J. Torres-Soto, E.A. Ashley, Multi-task deep learning for cardiac rhythm detection in wearable devices, NPJ Digital Medicine 3 (2020) 1–8.
[3] J.F. Thayer, S.S. Yamamoto, J.F. Brosschot, The relationship of autonomic imbalance, heart rate variability, and cardiovascular disease risk factors, Int. J. Cardiol. 141 (2010) 122–131.
[4] J. Bacevicius, Z. Abramikas, E. Dvinelis, D. Audzijoniene, M. Petrylaite, J. Marinskiene, J. Staigyte, A. Karuzas, V. Juknevicius, R. Jakaite, et al., High specificity wearable device with photoplethysmography and six-lead electrocardiography for atrial fibrillation detection challenged by frequent premature contractions: doubleCheck-AF, Frontiers in Cardiovascular Medicine 9 (2022) 869730.
[5] J. Ramesh, Z. Solatidehkordi, R. Aburukba, A. Sagahyroon, Atrial fibrillation classification with smart wearables using short-term heart rate variability and deep convolutional neural networks, Sensors 21 (2021) 7233.
[6] L. Zhu, V. Nathan, J. Kuang, J. Kim, R. Avram, J. Olgin, J. Gao, Atrial fibrillation detection and atrial fibrillation burden estimation via wearables, IEEE Journal of Biomedical and Health Informatics 26 (2021) 2063–2074.
[7] J.M. Kwon, Y.Y. Jo, S.Y. Lee, S. Kang, S.Y. Lim, M.S. Lee, K.H. Kim, Artificial intelligence-enhanced smartwatch ECG for heart failure-reduced ejection fraction detection by generating 12-lead ECG, Diagnostics 12 (2022) 654.
[8] M.V. Perez, K.W. Mahaffey, H. Hedlin, J.S. Rumsfeld, A. Garcia, T. Ferris, V. Balasubramanian, A.M. Russo, A. Rajmane, L. Cheung, et al., Large-scale assessment of a smartwatch to identify atrial fibrillation, N. Engl. J. Med. 381 (2019) 1909–1917.
[9] S.A. Lubitz, A. Faranesh, C. Selvaggi, S. Atlas, D.D. McManus, D.E. Singer, S. Pagoto, A. Pantelopoulos, A. Foulkes, Detection of atrial fibrillation in a large population using wearable devices: the Fitbit Heart Study, in: Proceedings of the Circulation. Lippincott Williams & Wilkins Two Commerce Sq, 2001 Market St, Philadelphia, vol. 144, 2021, pp. E570–E571.
[10] P. Badertscher, M. Lischer, D. Mannhart, S. Knecht, C. Isenegger, J.D.F. de Lavallaz, B. Schaer, S. Osswald, M. Kühne, C. Sticherling, Clinical validation of a novel smartwatch for automated detection of atrial fibrillation, Heart Rhythm O2 3 (2022) 208–210.
[11] Y. Guo, H. Wang, H. Zhang, T. Liu, L. Li, L. Liu, M. Chen, Y. Chen, G.Y. Lip, Photoplethysmography-based machine learning approaches for atrial fibrillation prediction: a report from the huawei heart study, JACC (J. Am. Coll. Cardiol.): Asia 1 (2021) 399–408.
[12] D. Hiraoka, T. Inui, E. Kawakami, M. Oya, A. Tsuji, K. Honma, Y. Kawasaki, Y. Ozawa, Y. Shiko, H. Ueda, et al., Diagnosis of atrial fibrillation using machine learning with wearable devices after cardiac surgery: algorithm development study, JMIR Formative Research 6 (2022) e35396.
[13] J. Wasserlauf, C. You, R. Patel, A. Valys, D. Albert, R. Passman, Smartwatch performance for the detection and quantification of atrial fibrillation, Circulation: Arrhythmia and Electrophysiology 12 (2019) e006834.
[14] G.H. Tison, J.M. Sanchez, B. Ballinger, A. Singh, J.E. Olgin, M.J. Pletcher, E. Vittinghoff, E.S. Lee, S.M. Fan, R.A. Gladstone, et al., Passive detection of atrial fibrillation using a commercially available smartwatch, JAMA Cardiology 3 (2018) 409–416.
[15] C.R. Cole, E.H. Blackstone, F.J. Pashkow, C.E. Snader, M.S. Lauer, Heart-rate recovery immediately after exercise as a predictor of mortality, N. Engl. J. Med. 341 (1999) 1351–1357.
[16] S. Nanas, M. Anastasiou-Nana, S. Dimopoulos, D. Sakellariou, G. Alexopoulos, S. Kapsimalakou, P. Papazoglou, E. Tsolakis, O. Papazachou, C. Roussos, et al., Early heart rate recovery after exercise predicts mortality in patients with chronic heart failure, Int. J. Cardiol. 110 (2006) 393–400.
[17] S.I. Nissinen, T.H. Mäkikallio, T. Seppänen, J.M. Tapanainen, M. Salo, M.P. Tulppo, H.V. Huikuri, Heart rate recovery after exercise as a predictor of mortality among survivors of acute myocardial infarction, Am. J. Cardiol. 91 (2003) 711–717.
[18] A.A.M. Duarte, C. Mostarda, M.C. Irigoyen, K. Rigatto, A single dose of dark chocolate increases parasympathetic modulation and heart rate variability in healthy subjects, Rev. Nutr. 29 (2016) 765–773.
[19] K.A. Sauder, E.R. Johnston, A.C. Skulas-Ray, T.S. Campbell, S.G. West, Effect of meal content on heart rate variability and cardiovascular reactivity to mental stress, Psychophysiology 49 (2012) 470–477.
[20] W. Hu, X. Jin, P. Zhang, Q. Yu, G. Yin, Y. Lu, H. Xiao, Y. Chen, D. Zhang, Deceleration and acceleration capacities of heart rate associated with heart failure with high discriminating performance, Sci. Rep. 6 (2016) 23617.
[21] A. Naseri, D. Tax, P. van der Harst, M. Reinders, I. van der Bilt, Data-efficient machine learning methods in the ME-TIME study: rationale and design of a longitudinal study to detect atrial fibrillation and heart failure from wearables, Cardiovascular Digital Health Journal 4 (2023) 165–172.
[22] E. Mejía-Mejía, J.M. May, P.A. Kyriacou, Effect of filtering of photoplethysmography signals in pulse rate variability analysis, in: Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2021, pp. 5500–5503.
[23] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al., SciPy 1.0: fundamental algorithms for scientific computing in Python, Nat. Methods 17 (2020) 261–272.
[24] L. Pardo, Statistical Inference Based on Divergence Measures, CRC press, 2018.
[25] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, F. Huang, A tutorial on energy-based learning, Predicting Structured Data 1 (2006).
[26] D. Andreini, G. Pontone, A.L. Bartorelli, P. Agostoni, S. Mushtaq, E. Bertella, D. Trabattoni, G. Cattadori, S. Cortinovis, A. Annoni, et al., Sixty-four–slice multidetector computed tomography: an accurate imaging modality for the evaluation of coronary arteries in dilated cardiomyopathy of unknown etiology, Circulation: Cardiovascular Imaging 2 (2009) 199–205.
[27] V.W. Anelli, T. Di Noia, E. Di Sciascio, C. Pomo, A. Ragone, On the discriminative power of hyper-parameters in cross-validation and how to choose them, in: Proceedings of the 13th ACM Conference on Recommender Systems, 2019, pp. 447–451.