NORDIC MACHINE INTELLIGENCE

Nordic Machine Intelligence, vol. 3, pp. 1–12, 2023
Received 07 Feb 2023 / published 23 May 2023
https://doi.org/10.2478/nmi.10140

Recurrent Neural Networks for Artifact Correction in HRV Data During Physical Exercise

Jakob Svane [1 2], Tomasz Wiktorski [2], Stein Ørn [3] and Trygve Christian Eftestøl [2]

1. *E-mail any correspondence to: jakob.svane@uis.no*
2. *Department of Electrical Engineering and Computer Science, University of Stavanger, Norway*
3. *Division of Cardiology, Stavanger University Hospital, Norway*

Abstract

In this paper, we propose the use of recurrent neural networks (RNNs) for artifact correction and analysis of heart rate variability (HRV) data. HRV can be a valuable metric for determining the function of the heart and the autonomic nervous system. When measured during exercise, motion artifacts present a significant challenge. Several methods for artifact correction have previously been proposed, none of them applying machine learning, and each presenting some limitations regarding an accurate representation of HRV metrics. RNNs offer the ability to capture patterns that might otherwise not be detected, yielding predictions where no prior physiological assumptions are needed.

A hyperparameter search has been carried out to determine the best network configuration and the most important hyperparameters. The approach was tested on two extensive multi-subject data sets, one from a recreational bicycle race and the other from a laboratory experiment. The results demonstrate that RNNs outperform by order of magnitude existing methods with respect to the calculation of derived HRV metrics. However, they are not able to accurately fill in individual missing RR intervals in sequence. Future research should pursue improvements in the prediction of RR interval lengths and reduction in necessary training data.

Keywords: HRV, heart rate variability, artifact, correction, RNN, recurrent neural network

Introduction

Heart rate (HR) variability (HRV) is the variation in time difference between successive heartbeats. The electrical signals that control the contraction of the various components of the heart are referred to as the PQRST-complex, where the QRS-complex is responsible for ventricular depolarization [1]. The R-wave coincides with the contraction of the left ventricle, commonly known as the heartbeat. By measuring the distance between consecutive R-waves, we get the RR interval length, i.e., the time between heartbeats. These electrical signals can be measured by using an electrocardiogram (ECG), which registers the heart's electrical activity. Since the RR interval length varies, the electrical signal is sampled irregularly, making HRV data irregular. Fig. 1 shows three typical PQRST-complexes, and Fig. 2 displays a plot of typical RR interval sampling.

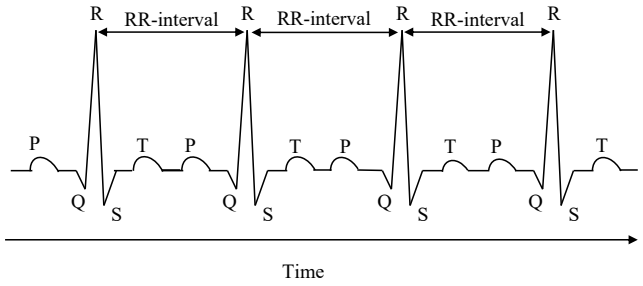

Figure 1: Three PQRST-complexes displaying the typical pattern of electrical activity to the heart. This figure is for illustration purposes only, and does not represent real data.

The heart is controlled by the autonomic nervous system (ANS), with the sympathetic branch increasing HR, and the parasympathetic branch lowering HR. HRV is higher during parasympathetic activity and lower during sympathetic activity [2]. Increases in HR can occur directly from sympathetic activation, but also from the interaction of levels of expression of vagal and sympathetic activation. Notably, changes in HR level can be as important as changes in HRV for ANS assessment. An important component of the beat-to-

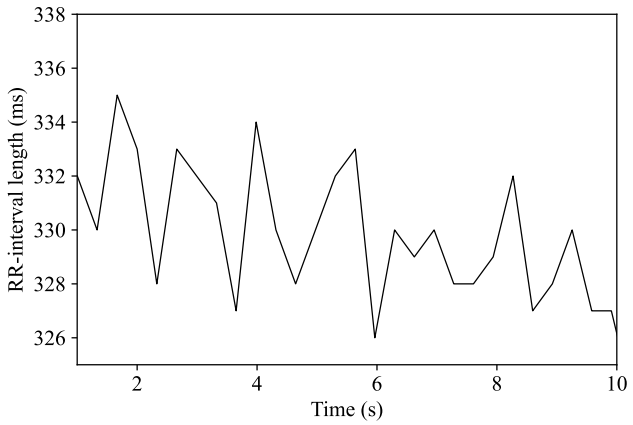

Figure 2: Example of RR interval length plotted over time during exercise. It should be noted that the RR interval length here is substantially shortened compared to during non-strenuous activity. The data are gathered from a subject from the NEEDED study.

beat variability is respiratory sinus arrhythmia (RSA). RSA is the variation in HR within a respiratory cycle, with HR slightly increasing during inspiration, and decreasing during expiration [3]. During rest, the amplitude of HRV can be considered an expression of vagal activity. Thus, HRV can be used as a noninvasive assessment of the autonomic nervous system regulation, as well as an indicator of underlying cardiovascular diseases [4]. Importantly, assessment of respiration should be considered to validate RSA for vagal influence (i.e., when reproducing circumstances for HRV measurements during rest, breathing frequency should be identical).

The most common metrics regarding HRV in the time domain include the root mean square of successive differences (RMSSD) and the standard deviation of all normal-normal intervals (SDNN), where normal-normal intervals are error-free RR intervals. In addition, three frequency bands in the power spectrum, as well as the sum of these, are commonly used: very low frequency (VLF), 0Hz–0.04Hz; low frequency (LF), 0.04Hz–0.15Hz; high frequency (HF), 0.15Hz–0.4Hz; and total power (TP). To calculate the frequency domain metrics, the HRV-data are usually interpolated and resampled evenly, and then the Fast Fourier Transform (FFT) is applied [5]. The Lomb-Scargle-Periodogram method has also been suggested for ultra-short-term recordings [6].

Objective

Data artifacts present a significant issue for the analysis of HRV data, as they produce unrealistic RR interval lengths and consequently erroneous HRV metrics. Measurements taken during exercise are particularly prone to motion artifacts, making reliable artifact correction methods crucial. The objective of this paper is to test the performance of recurrent neural networks as an artifact correction method on HRV data measured during exercise. It is important to note that we are interested in gap-filling, not prediction. Detection of artifacts will not be included in this paper.

Background

NEEDED Project

HRV data may differ between individuals with and without heart disease [7]. The North Sea Race Endurance Study Research program (NEEDED) [8] works to identify asymptomatic heart disease based upon alterations in biomarker response to physical exercise [9]. The program was initiated in 2012 and has included more than 1100 presumably healthy asymptomatic individuals in several clinical studies. The current phase of the NEEDED research program develops new assessment methods using new exercise protocols, biomarkers and HRV data, both at rest and during exercise, to determine the relationship between these markers and heart disease in different populations. Measurements in relation to these protocols require HRV data derived from shorter duration than traditional HRV studies, which renders these protocols more susceptible to be influenced by artifacts and therefore require accurate artifact correction methods.

Artificial Intelligence in Wearables

With advances in wearable sensor technology, the possibility of incorporating artificial intelligence (AI) algorithms as a part of data handling in wearable sensors increases [10]. Coutts et al. [11] applied deep learning on HRV data from wearable devices for the prediction of mental and general health, demonstrating the potential of combining artificial intelligence and wearable tracking. It has been proposed that most machine learning based technologies that detect cardiovascular outcomes using wearables are not operational due to failing to acquire proper realistic data [12]. The same paper also suggests more frequent application of sequential neural networks in data processing. Challenges with regard to data collection and data processing when using AI in wearables have also been highlighted [13].

HRV Data Artifacts

Typically, HRV has been measured in controlled settings to assess parasympathetic activity using ECG. However, with the development of HR chest straps and photoplethysmography (PPG), it is now also possible to measure HRV during movement and exercise [10]. Analyzing HRV during exercise might provide insight that is not available during rest, as the heart and cardiovascular system is placed under more strain [14]. A significant challenge with HRV monitoring during exercise is the introduction of movement-related artifacts that distort the signal, causing changes to the calculated metrics. It has been shown that a single artifact on a 4-minute reading could increase RMSSD by as much as 413% [15]. Cajal et al. [16] found that in order to obtain estimations with an error less than 20% for both time domain and frequency

domain metrics, segments with more than 25% of missing beats should be discarded.

Previous Work

Common methods for artifact correction include deletion and linear, cubic and spline interpolation [17]. Although widely used, these methods have several drawbacks. Deletion causes discontinuity and loss of samples, which affect spectral analysis to a large extent, whereas interpolation induces bias and produces an unrealistic beat-to-beat variability [17]. Other methods for artifact correction during resting conditions [16, 18], as well as for during exercise [19], have also been proposed. Despite working well for shorter gaps, they are less accurate for larger gaps, causing larger errors in the HRV metrics. Moreover, it has been shown that the application of different artifact correction methods can have a significant impact on HRV metrics [20]. Neural networks are widely used in time series data handling and prediction [21], including for ECG signal handling [22], but have not been used for raw RR interval data artifact correction, to our knowledge.

Detection

Artifact detection is often handled by threshold filtering and visual inspection [23]. Azar et al. [24] used a combination of a convolutional neural network and a recurrent neural network autoencoder for PPG artifact filtering, achieving 90% precision and 95% recall. Zubair et al. [22] applied a recurrent neural network and a deep neural system for removing motion artifacts in ECG signals. In many cases, detection is also carried out manually by an experienced human reviewer. For complete automatic HRV artifact handling, both detection and correction are needed. Nevertheless, detection will not be a part of the current paper, as we will rather focus on correction.

Ultra-short-term Recordings

When measuring HRV in the short-term, recordings of 5 minutes are recommended [5]. The use of ultra-short-term (UST) analysis as a surrogate of standard 5-minute analysis has also been investigated [25, 26], as it provides a more convenient and accessible method for measurement. Exercise causes a transient physiological state, thus imposing the need for reliable UST analysis. Baek et al. [25] explored the reliability of UST analysis, showing insignificant differences in RMSSD calculated using 30-second interval length, SDNN from 240-second interval length, LF from 90-second interval length, HF from 20-second interval length, LF from 90-second interval length, and VLF from 270-second interval length, compared with the same HRV variables calculated using standard 5-minute interval data. For UST recordings, the effect of artifacts is magnified, making artifact handling essential [27]. Importantly, caution must be exercised when assessing ANS regulation based on UST recordings.

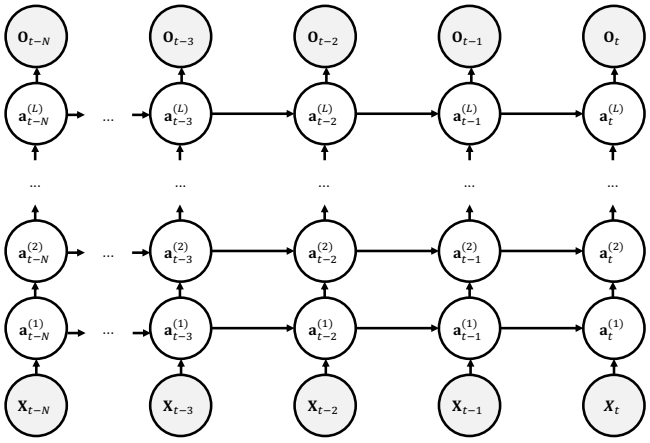

Figure 3: Architecture of a deep recurrent neural network. The figure shows how the input, $\mathbf{X}$, is fed to the hidden layers, $\mathbf{a}$, which in turn lead to the output, $\mathbf{O}$. The subscript denotes the time step, and the superscript denotes the hidden layer. Thus, in this figure, there are N+1 time steps and L hidden layers.

Due to the effect of respiration on HRV, a low breathing rate may yield only a few cycles of variability for a measurement of ultra short duration. Whether a true cyclicity in frequency domain metrics can be inferred from such a minimal amount of cycles must therefore be considered when interpreting the results.

Recurrent Neural Networks

Recurrent neural networks (RNNs) are network architectures that are designed to work with sequential data. RNNs can be used to make predictions on sequential data based on previous data points, which in turn can work as a method for imputing missing values. Figure 3 illustrates a typical network structure for RNNs.

Based on Figure 3, the equations constituting a RNN can be described. For any hidden layer $\mathbf{a}_t^{(L)}$, where the superscript denotes the layer and the subscript denotes the time step, its value can be calculated as

$$\mathbf{a}_t^{(L)} = f(\mathbf{W}_a \mathbf{a}_t^{(L-1)} + \mathbf{W}_t \mathbf{a}_{t-1}^{(L)} + \mathbf{b}), \tag{1}$$

where f is an activation function such as sigmoid or tanh, $\mathbf{W}_a$ and $\mathbf{W}_t$ are the weights associated with the previous layer and previous time steps, respectively, and $\mathbf{b}$ is the sum of the bias associated with the previous layer and previous time step. The value for the first layer at any time step (second to bottom row in Figure 3) is

$$\mathbf{a}_t^{(1)} = f(\mathbf{W}_t \mathbf{a}_{t-1}^{(1)} + \mathbf{W}_x \mathbf{X}_t + \mathbf{b}), \tag{2}$$

with $\mathbf{X}_t$ being the input at time step t, and $\mathbf{W}_x$ being the weight associated with the input. At the first time step, t_N, where N is the number of previous time steps considered, the value of the first hidden layer is only based on the input. Knowing Equations (1) and (2), the output at time step t can be calculated as

$$\mathbf{O}_t = f(\mathbf{W}_t \mathbf{a}_t^{(L)} + \mathbf{b}). \quad (3)$$

Considering these equations, we see that the output at time step t is not only based on its own input, but also the values of the N previous time steps. To apply the model, the weights are randomly initialized and all values are calculated in the forward propagation, before the error is calculated and the backwards propagation updates the weights. A thorough explanation of various design possibilities and applications of RNNs can be found in [28].

Two of the most popular RNNs are Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) networks. Staudemeyer and Morris [29] give a detailed and clear explanation of how LSTM works, and Chung et al. [30] detail how GRU works, as well as the difference between the two. Greff et al. [31] performed a large-scale analysis of various LSTM variants, showing that no variants improve significantly upon the standard LSTM, but that GRU simplifies the model without compromising performance significantly. We have therefore limited this paper by only considering these two architectures.

Data Set

The data sets inspected in this paper consist of HRV data from two different databases. From the North Sea Endurance Exercise Study (NEEDED), HRV data from 15 different individuals were collected with a Garmin HRM chest strap during a bicycle race. For each participant, the HRV data from the whole race were visually inspected, and the longest section of data with no visible artifacts and no missing samples was extracted. Fig. 4 shows a segment of raw HRV data containing artifacts, and an artifact-free section within the original segment. The average length of the sections was 5 minutes and 2 seconds, with the shortest being 3 minutes and 10 seconds and the longest being 8 minutes and 1 second. The sections were extracted from random locations in the race for different participants and include both HR acceleration and deceleration. Fig. 5 shows the altitude profile and HR for a section of the race for one of the participants.

Data from a separate laboratory experiment using 12 volunteers were also used. These were the Physionet simultaneous physiological measurements [32, 33]. The data from Physionet were collected at four different physiological loads of 5 minutes each using a Polar RS800 Multi chest strap. Only the 5 minute segment corresponding to uphill walking on the treadmill (15% track inclination, 1.2 m/s) was extracted for use in this paper. One participant from the Physionet database was excluded, due to apparent data errors during the 5-minute segment (participant 5).

By utilizing the data from both NEEDED and Physionet, the method can be tested on two completely independent data sets. NEEDED provides data from the field, at different work-loads and external conditions, whereas

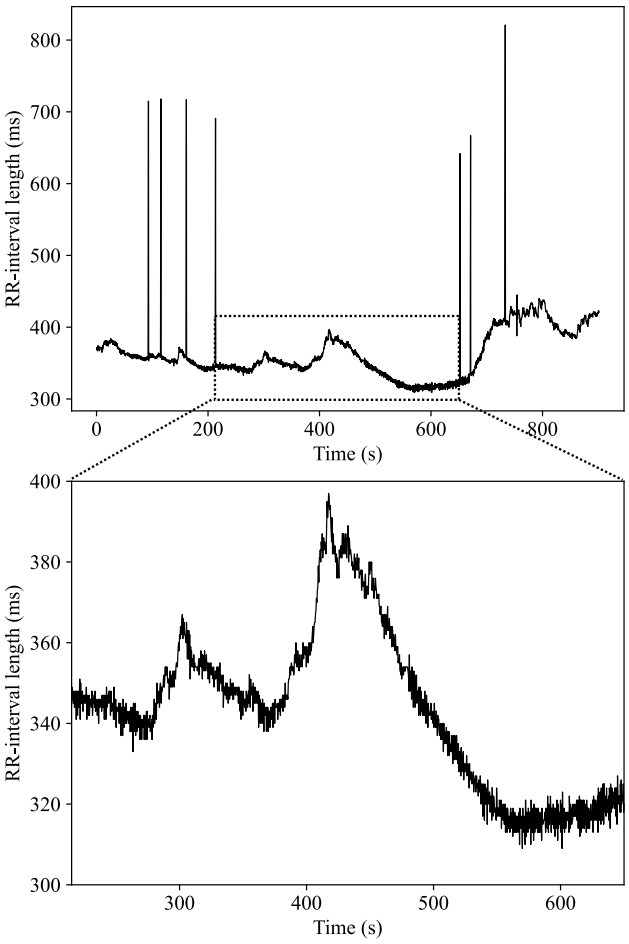

Figure 4: Two segments of RR interval length plotted over time. One containing artifacts (top), and the other a section of the first without artifacts (bottom). The data are gathered from a subject from the NEEDED study.

Physionet provides data from a more controlled laboratory setting. There is also the advantage that the NEEDED data was collected during a bike race and the Physionet data were collected while uphill walking on a treadmill, possibly validating the current method for both modalities.

In both data sets, the data consist of the RR interval lengths in seconds, exemplified in Figure 2. The data are irregularly sampled, as each data point is sampled at the detection of the R-wave by the HR chest strap.

Materials and methods

Hyperparameter Search

When building and training a RNN, there are several hyperparameters that can be adjusted to configure the network architecture and training configuration. In addition to hyperparameters such as batch size, dropout, epochs, hidden layers, learning rate and units per layer in traditional non-temporal neural networks, RNNs also include sampling rate and sequence length. Sequence length is the number of time steps that are used as input

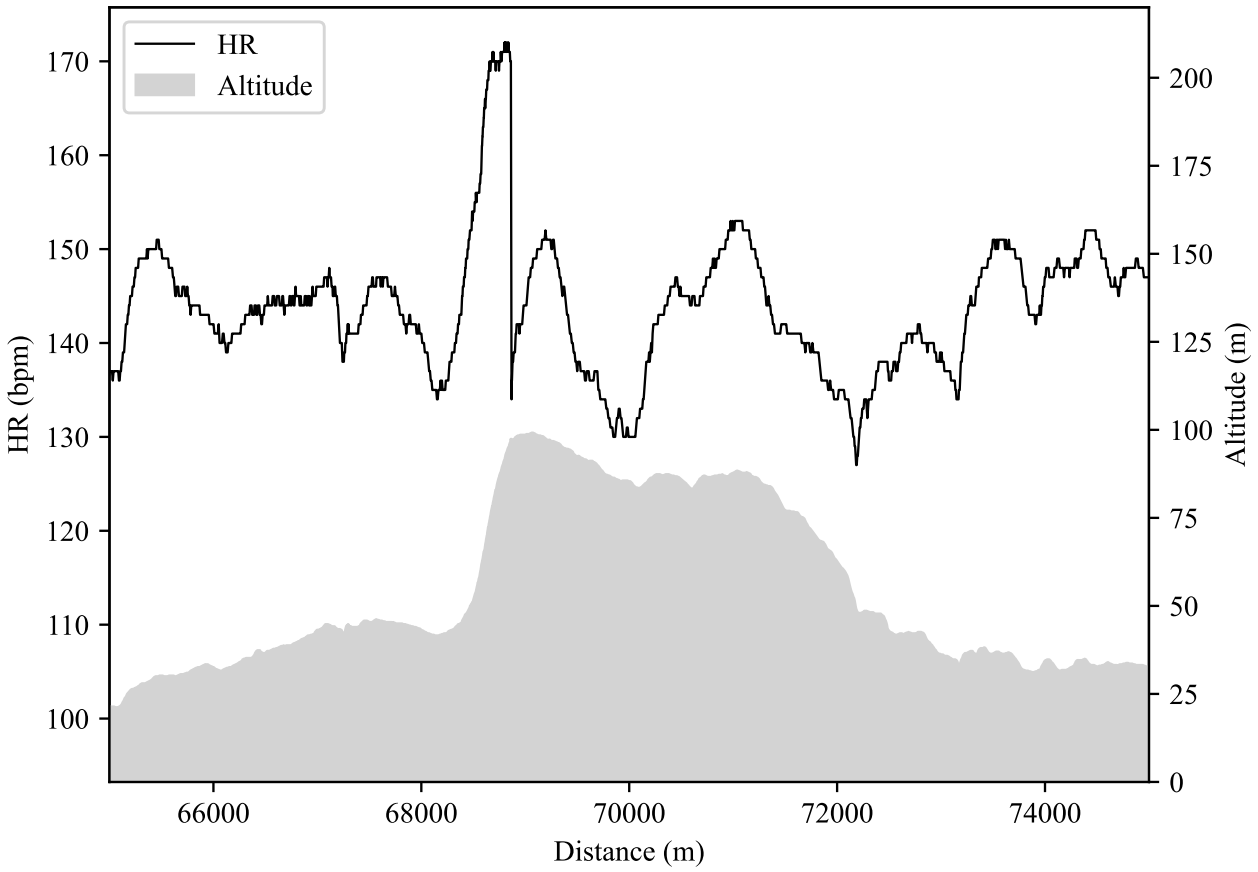

Figure 5: Altitude profile (grey shaded area) and heart rate for example participant from NEEDED during section of the race.

for the prediction of the following value (N in Figure 3), whereas sampling rate is simply the rate of samples used in the sequence length. With the objective of the paper being gap filling, bidirectional layers may also be utilized. Bidirectional layers allow the RNN to be trained in the positive and negative time direction simultaneously. In other words, both previous and future data points are used as input to predict the missing values, in many cases improving results [34].

Moreover, two different methods for prediction can be used. In rolling prediction, one value is predicted at a time, and then used as input when predicting the next value. In this case, the RNN is trained for only one step ahead, but the process is repeated for the desired number of predicted steps. The loss is calculated as the error between the predicted value and the true value. The error calculation depends on which loss function is chosen. In single-shot prediction, the model is trained for the desired number of steps ahead, with all values predicted at once. The loss is calculated as the mean of the error between each element in the output vector and the true values.

To test the various configurations of RNN as an artifact correction method, a hyperparameter search was carried out using the sweep tool from WandB [35]. First, a larger search was performed for three participants' data from NEEDED to get a sense of the influence and importance of the various hyperparameters. Subsequently, the search was narrowed for the remaining participants' data to save computational cost. We tested three different loss functions: root mean square error (L_{RMSE}), error in RMSSD (L_{RMSSD}), and error in dynamic time warping (L_{DTW}). The loss function L_{RMSE} is defined as the Euclidean distance between the predicted and true values, whereas L_{RMSSD} was calculated as the difference in the true RMSSD and the predicted RMSSD, and L_{DTW} applied dynamic time warping [36] between the true and predicted values. Further, we assessed both LSTM and GRU, as well as whether applying differencing one or two times beforehand would improve the results by removing trends. Bidirectional layers were also tested. These layers include data points following immediately after the gap as input, and can be applied since we are interested in gap-filling and not prediction. The full range of the hyperparameter values are shown in Table 1.

The work in this paper was carried out using Keras 2.9.0 [37] in Python 3.9.7.

Table 1: All the parameters tested in the hyperparameter search and the range of values for each.

	Range of values		
Batch size	10-100%		
Bidirectional Layers	No	Yes	
Differencing	No	Yes	Two times
Dropout	0 - 20%		
Epochs	30 - 80		
Hidden layers	1 - 8		
Learning rate	0.001 - 0.1		
Loss function	L_{RMSE}	L_{DTW}	L_{RMSSD}
Network type	GRU	LSTM	
Sampling rate	1 - 2		
Sequence length	1 - 100		
Units per layer	32 - 1024		

Prediction

Each participant's data were split in three sections: a training set (50%), a validation set (20%) and a test set (30%). Next, a gap of length 1, 3, 7 and 10 was introduced in the test set by removing successive values. At a random location, one gap was introduced in each test set (as long as the number of points before were equal to or bigger than the sequence length). Finally, the missing values were replaced by the RNNs' predictions, using either rolling prediction or single-shot prediction. This way the true HRV metrics can be compared to those of the gap-filling approach. When using single-shot prediction, only a gap size of 10 is evaluated since we are mostly interested in longer gaps, and one would need to retrain the model for each gap size. The hyperparameter search was then carried out on the training and validation sets, and the performance of each configuration tested on the test data. The results from the RNNs were then compared to the results from both cubic spline interpolation and deletion of the artifact, as a baseline for performance. It should be noted that the method assumes that the number of missing beats are known, which is not always true. However, this issue will not be addressed in this paper, as we consider it a part of the preceding artifact detection.

In this paper, the models are trained on several data sets from two different databases. However, in a real-life setting, there might only be one data set. In this instance, the application of k-fold cross-validation (CV) might better fit the purpose. With k-fold CV, the data set is split into a training set and a test set, and the training set is divided into k groups, where the model is trained on $k-1$ groups, and the last group is used as validation set. This is then repeated for all k groups. This requires more computational power, but can be very useful for finding optimal hyperparameters on a smaller data set. Additionally, using k-fold CV, the uncertainty of the model can be estimated, indicating the credibility of the model.

Evaluation

The performance of the RNN models is evaluated by calculating the relative error in RMSSD, SDNN, VLF, LF, HF and TP from the predictions made by the RNN models. It is then compared to cubic spline interpolation, as it is the most commonly used artifact correction method for HRV data [17]. It is also compared to deletion, i.e., simply removing the data points where the artifacts occur. The frequency domain metrics are calculated by resampling by interpolation and then applying FFT. A visual comparison of the predicted and true RR interval lengths is also performed. To evaluate the importance of each hyperparameter, the parameter-importance function (PIF) from WandB is applied. The PIF works by training a random forest with the hyperparameters as input and the metric as the target output. Based on the parameter's position in the decision trees in the random forest algorithm, the parameter is given a score ranging from 0 to 1, so that the sum of all parameters equals 1 (i.e., a score of 1 means that the parameter is the only relevant parameter, whereas a score of 0 means that it is not relevant at all). This way, the importance of each hyperparameter on each of the HRV metrics can be assessed. It is important to note that the PIF is affected by the number of values tested, and should not be considered a definite result, but rather be used as an indication for which parameters should be tuned. Thus, the PIF was inspected only for three participants in the NEEDED study, but with a large range of values for each parameter. The results were used to select the most important parameters for the hyperparameter search, allowing us to narrow the search on the other participants to reduce computational costs.

Results

Hyperparameter Search

The hyperparameter search showed that for all metrics, the worst-performing models were those with the L_{RMSSD} loss function, and the best were usually those with one-differenced data sets. There were no clear indications on whether LSTM or GRU performed better, nor which loss function gave the best results between L_{RMSE} and L_{DTW}. Models using L_{DTW} were significantly slower in training, thus L_{RMSE} seems the most practical loss function. The PIF for the three first hyperparameter searches is summarized in Fig. 6, showing that the most important hyperparameter for each metric was sequence length, followed by the number of hidden layers and units. There were no clear indications of which values these parameters should have. Although learning rate, batch size, number of epochs, dropout and sampling rate affected the results as well, they were not included in the hyperparameter searches on the remaining data sets, as they were less significant. That is not to say that

tuning these hyperparameters as well would not improve the model, but rather that they were omitted here to reduce computational cost.

To demonstrate the impact of these three hyperparameters on the models, Figures 7-9 show how the error in HF changes for different values of sequence length, hidden layers and units for three different participants from NEEDED. All other hyperparameters are kept the same. Error in HF is chosen for these figures since it is one of the metrics with the biggest error using common approaches such as cubic interpolation or deletion, but it could have been any of the other metrics.

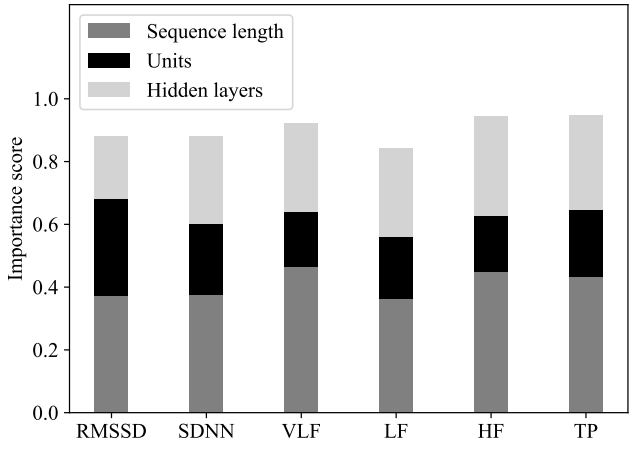

Figure 6: Importance of the three most important parameters for each metric using rolling prediction, for the three extended hyperparameter searches. RMSSD=root mean square of successive differences; SDNN=standard deviation of NN-intervals; VLF=very low frequency; LF=low frequency; HF=high frequency; TP=total power.

HRV Metrics

Relative errors in the HRV metrics for the RNNs using rolling prediction, cubic spline interpolation and deletion on the NEEDED data for varying gap sizes are shown in Table 2. The RNNs outperform the cubic spline interpolation and deletion method for all metrics, across all gap sizes evaluated. In particular, the error in the frequency domain metrics was improved upon the most, compared to cubic spline and deletion. Single-shot prediction yields similar results, as can be seen in Table 3. It can also be seen in Table 3 that the RNNs on a gap size of ten outperform the method from Królak et al. [19] on a gap size of seven, for the same data. It should be noted that there is no significant difference in relative error when applying bidirectional layers as compared to unidirectional layers.

RR Interval Length

The shapes of the predicted values for the RR interval lengths were visually not similar to the true values when using rolling prediction. Fig. 10 shows the predicted

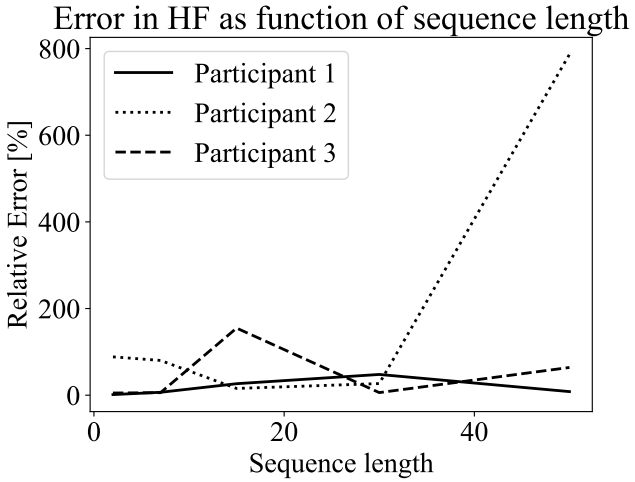

Figure 7: The error in HF for different sequence lengths, with all other hyperparameters the same, on three different participants from NEEDED. HF = high frequency.

values for RNNs with rolling prediction and cubic spline for a gap of size 10 for two participants from Physionet and NEEDED. Notably, the predicted values tend to either steadily increase or decrease from the first predicted value to the last predicted value. On the other hand, single-shot prediction seems to yield slightly more appropriate results. Fig. 11 shows the predicted values using single-shot prediction with unidirectional layers, whereas Fig. 12 shows the predicted values using bidirectional layers.

Evidently, the single-shot prediction better captures the variability of the data than the rolling prediction, although it is not able to precisely predict the true values. The bidirectional layers seem to improve the results compared to the unidirectional layers. It should be noted that choosing which model's prediction to use is not straightforward. The values chosen in the figures are those of the model with the smallest validation loss. However this approach will not necessarily always give the best results. Figure 13 shows two participants from Physionet and NEEDED in which using the lowest validation loss yields poor results. Also, the best performing model varies between subjects.

Discussion

The objective of this paper was to test the performance of RNNs as a method for HRV data artifact correction. The data were retrieved from 15 different individuals during a bicycle race (NEEDED) and from 12 individuals during a treadmill experiment (Physionet), both measured with HR chest straps.

Hyperparameter Search

A hyperparameter search was carried out to achieve the best results with the RNNs. Sequence length, number of units per layer and number of hidden layers seem to be the

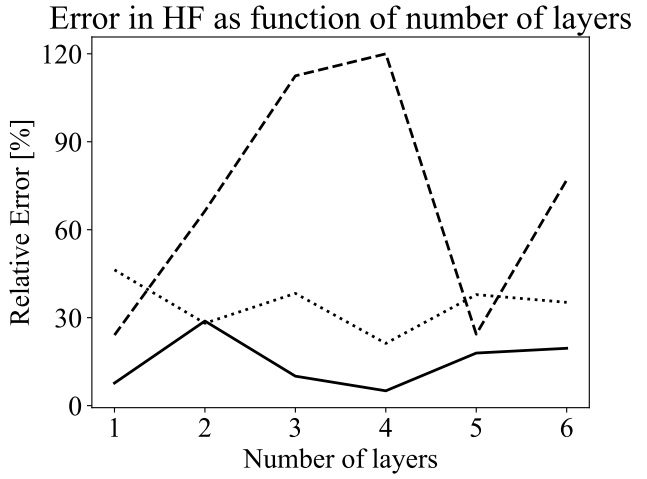

Figure 8: The error in HF for different numbers of layers, with all other hyperparameters the same, on three different participants from NEEDED. HF = high frequency.

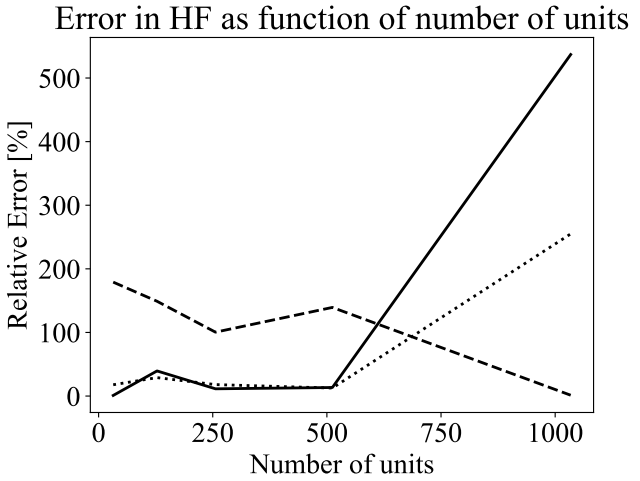

Figure 9: The error in HF for different numbers of units, with all other hyperparameters the same, on three different participants from NEEDED. HF = high frequency.

Table 2: Relative mean error for typical HRV metrics for various gap sizes after applying the best RNN model with rolling prediction, cubic spline interpolation and deletion on the data from NEEDED. RMSSD=root mean square of successive differences; SDNN=standard deviation of NN-intervals; VLF=very low frequency; LF=low frequency; HF=high frequency; TP=total power; RNN=recurrent neural network; CUBIC=cubic spline interpolation.

Gap size	Method	RMSSD	SDNN	VLF	LF	HF	TP
1	RNN	0.15%	0.01%	0.29%	2.02%	0.35%	0.12%
	CUBIC	0.45%	0.09%	1.24%	8.35%	4.75%	0.56%
	DELETION	0.59%	0.37%	8.61%	18.32%	7.64%	4.64%
3	RNN	0.14%	0.01%	0.57%	3.21%	2.20%	0.17%
	CUBIC	0.94%	0.12%	1.53%	13.87%	8.74%	1.18%
	DELETION	0.89%	0.59%	9.84%	27.69%	15.61%	7.54%
7	RNN	0.26%	0.07%	0.57%	5.79%	4.53%	0.24%
	CUBIC	1.56%	0.41%	7.02%	55.71%	38.94%	4.76%
	DELETION	1.57%	1.16%	17.33%	54.88%	15.36%	12.01%
10	RNN	0.31%	0.06%	1.26%	9.37%	3.59%	0.84%
	CUBIC	2.15%	1.02%	12.51%	12.78%	36.01%	8.86%
	DELETION	2.00%	1.71%	20.86%	73.39%	45.16%	16.87%

most important parameters to tune correctly, but there were no clear indications on which values gave the best performing models. Figures 7-9 also show the difficulty in choosing the optimal hyperparameters, as there is no clear trend in HF error, and the models act differently for different data sets. One-differencing as a preprocessing step improved the performance. There were no apparent differences in performance between GRU and LSTM, supporting the results from Greff et al. [31]. Bidirectional layers seemed to better capture the variability of the data than unidirectional layers, but impose the issue of needing continuous error-free data directly following the gap as well.

Table 3: Relative mean error for typical HRV metrics for a gap size of ten after applying the best RNN model with single-shot prediction, cubic spline interpolation and deletion on the data from Physionet. The results from Królak et al. [19] on the same data are also included, but for a gap size of seven. RMSSD=root mean square of successive differences; SDNN=standard deviation of NN-intervals; VLF=very low frequency; LF=low frequency; HF=high frequency; TP=total power; RNN=recurrent neural network; CUBIC=cubic spline interpolation; N.A=not available.

Gap size		RMSSD	SDNN	VLF	LF	HF	TP
7	Królak et al.	1.42%	0.85%	N.A	6.13%	4.74%	2.03%
10	RNN	0.62%	0.24%	1.96%	3.01%	1.85%	2.45%
	CUBIC	2.63%	1.98%	24.64%	76.29%	51.7%	30.23%
	DELETION	1.30%	1.44%	29.76%	26.86%	18.53%	18.12%

Performance

The best performing RNNs outperformed cubic spline interpolation and deletion for all gap sizes and all metrics. In particular, frequency domain metrics were significantly more accurate with RNNs. Improving the error in the frequency domain metrics is very beneficial, as it has been shown that the frequency domain is particularly sensitive to artifacts in HRV data [38, 39]. The RNNs gave better results than the method from Królak et al. [19] on the same data, even when the gap size was longer for the RNNs. Although the RNNs in this paper were not able to predict true RR interval lengths, HRV is usually expressed as a derived metric rather than by the timeseries in itself, thus proving the usefulness for RNNs as an artifact correction method.

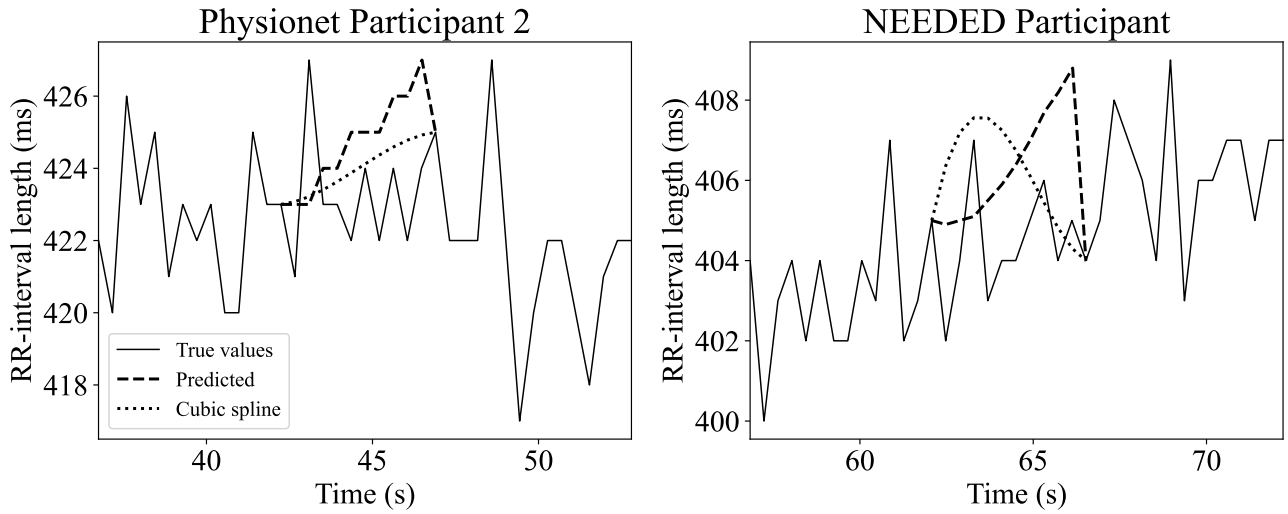

Figure 10: Comparison of the true RR interval lengths with RNN rolling prediction and cubic spline interpolation for gap size 10 for two different participants.

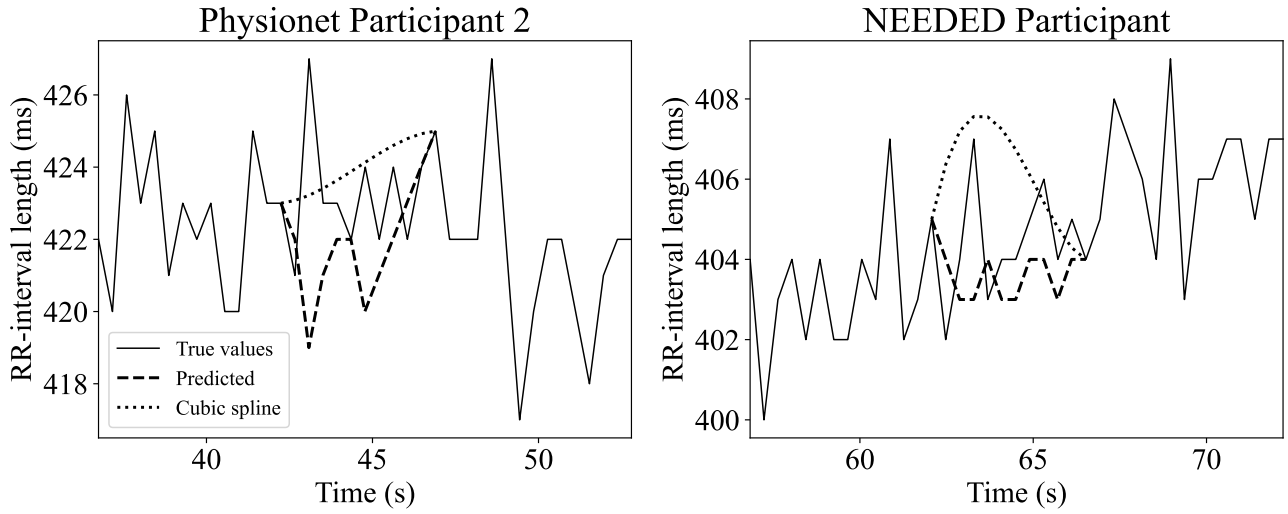

Figure 11: Comparison of the true RR interval lengths with RNN single-shot prediction and cubic spline interpolation for gap size 10 for two different participants.

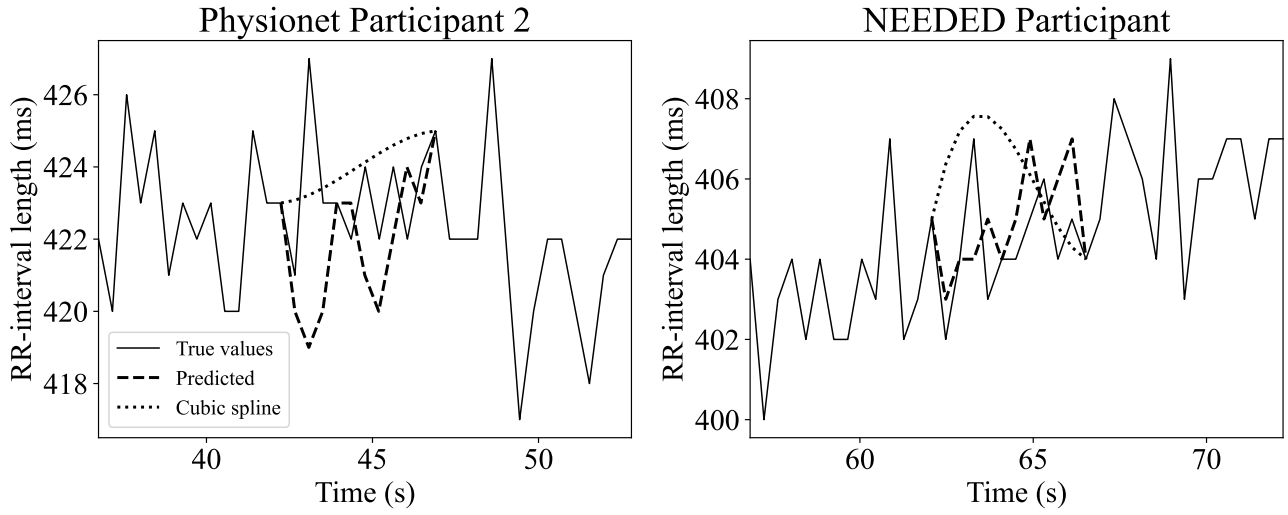

Figure 12: Comparison of the true RR interval lengths with RNN single-shot prediction and bidirectional layers and cubic spline interpolation for gap size 10 for two different participants.

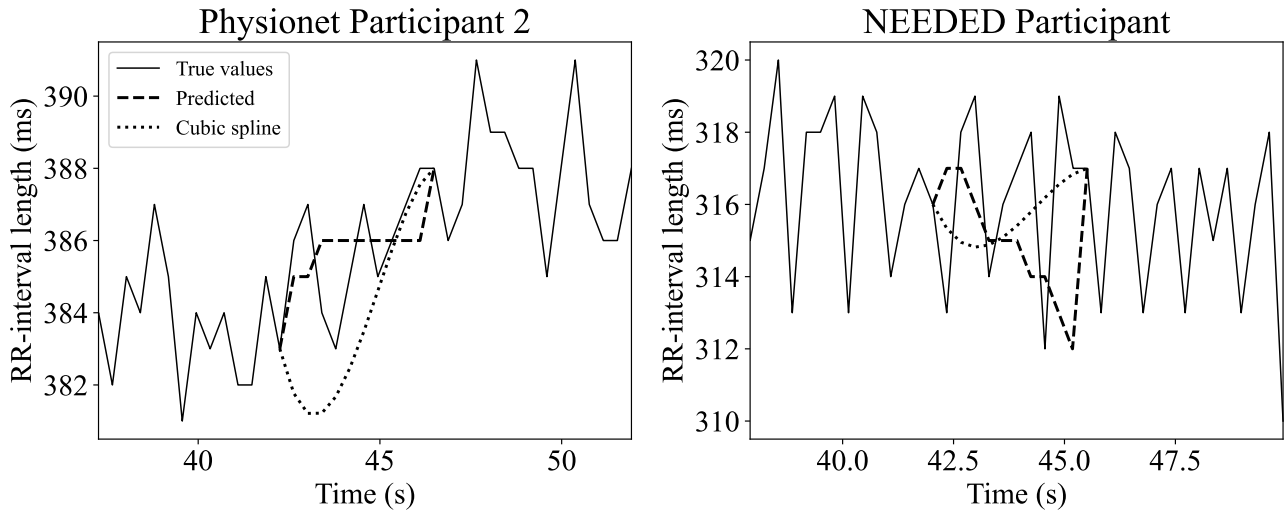

Figure 13: Example of poor predictions made by RNNs with single-shot prediction and bidirectional layers when choosing values based on the models' validation loss.

Limitations and Future Work

It is desirable to improve the prediction of RR interval lengths, since there are other HRV metrics not mentioned in this paper that also depend on error-free data [40]. Also, applying RNNs necessitates segments of continuous error-free training data, which is not always available [38]. In addition, training RNNs, and in particular performing hyperparameter searches, is computationally expensive and not always feasible. Choosing the best model is not straightforward, as a visual inspection of the predicted values shows that the model with the lowest validation loss will not always yield the best results, and that the best model depends on the data set. Hence, a network configuration that performs well across all data sets and metrics should be pursued. A possible solution to all of these problems might be to gather a large database of error-free data and applying transfer learning as a method for reducing the amount of individual training data and create a general model for all subjects [41].

Conflict of interest

Authors state no conflict of interest.

Resources

Our software implementation can be found at https://github.com/jakobsv97/HRV-RNN.

References

1. Scher A. Studies of the electrical activity of the ventricles and the origin of the QRS complex. Acta cardiologica 1995; 50:429–65
2. Dong JG. The role of heart rate variability in sports physiology. Experimental and therapeutic medicine 2016; 11:1531–6
3. Berntson GG, Thomas Bigger Jr J, Eckberg DL, Grossman P, Kaufmann PG, Malik M, Nagaraja HN, Porges SW, Saul JP, Stone PH, et al. Heart rate variability: origins, methods, and interpretive caveats. Psychophysiology 1997; 34:623–48
4. Kleiger RE, Miller JP, Bigger Jr JT, and Moss AJ. Decreased heart rate variability and its association with increased mortality after acute myocardial infarction. The American journal of cardiology 1987; 59:256–62
5. Camm AJ, Malik M, Bigger JT, Breithardt G, Cerutti S, Cohen RJ, Coumel P, Fallen EL, Kennedy HL, Kleiger RE, et al. Heart rate variability: standards of measurement, physiological interpretation and clinical use. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. 1996
6. Wehler D, Jelinek HF, Gronau A, Wessel N, Kraemer JF, Krones R, and Penzel T. Reliability of heart-rate-variability features derived from ultra-short ECG recordings and their validity in the assessment of cardiac autonomic neuropathy. Biomedical Signal Processing and Control 2021; 68:102651
7. Goldenberg I, Goldkorn R, Shlomo N, Einhorn M, Levitan J, Kuperstein R, Klempfner R, and Johnson B. Heart rate variability for risk assessment of myocardial ischemia in patients without known coronary artery disease: The HRV-DETECT (heart rate variability for the detection of myocardial ischemia) study. Journal of the American Heart Association 2019; 8:e014540
8. The North Sea Race Endurance Exercise Study (NEEDED). 2016. Available from: https : / / helse - stavanger . no / fag - og - forskning / forskning - i - helse - stavanger / the - north - sea - race - endurance - exercise - study-needed
9. Kleiven Ø, Omland T, Skadberg Ø, Melberg TH, Bjørkavoll-Bergseth MF, Auestad B, Bergseth R, Greve OJ, Aakre KM, and Ørn S. Occult obstructive coronary artery disease is associated with prolonged cardiac troponin elevation following strenuous exercise. European journal of preventive cardiology 2020; 27:1212–21
10. Ludwig M, Hoffmann K, Endler S, Asteroth A, and Wiemeyer J. Measurement, prediction, and control of individual heart rate responses to exercise—Basics and options for wearable devices. Frontiers in physiology 2018; 9:778
11. Coutts LV, Plans D, Brown AW, and Collomosse J. Deep learning with wearable based heart rate variability for prediction of mental and general health. Journal of Biomedical Informatics 2020; 112:103610
12. Jahfari AN, Tax D, Reinders M, Bilt I van der, et al. Machine Learning for Cardiovascular Outcomes From Wearable Data: Systematic Review From a Technology Readiness Level Point of View. JMIR medical informatics 2022; 10:e29434
13. Nahavandi D, Alizadehsani R, Khosravi A, and Acharya UR. Application of artificial intelligence in wearable devices: Opportunities and challenges. Computer Methods and Programs in Biomedicine 2022; 213:106541
14. Aubert AE, Seps B, and Beckers F. Heart rate variability in athletes. Sports medicine 2003; 33:889–919
15. Bourdillon N, Yazdani S, Vesin JM, Schmitt L, and Millet GP. RMSSD Is More Sensitive to Artifacts Than Frequency-Domain Parameters: Implication in Athletes' Monitoring. Journal of Sports Science and Medicine 2022; 21:260–6
16. Cajal D, Hernando D, Lázaro J, Laguna P, Gil E, and Bailón R. Effects of Missing Data on Heart Rate Variability Metrics. Sensors 2022; 22:5774
17. Giles DA and Draper N. Heart rate variability during exercise: a comparison of artefact correction methods. The Journal of Strength & Conditioning Research 2018; 32:726–35
18. Rincon Soler AI, Silva LEV, Fazan Jr R, and Murta Jr LO. The impact of artifact correction methods of RR series on heart rate variability parameters. Journal of Applied Physiology 2018; 124:646–52
19. Królak A, Wiktorski T, Bjørkavoll-Bergseth MF, and Ørn S. Artifact correction in short-term hrv during strenuous physical exercise. Sensors 2020; 20:6372
20. Alcantara JM, Plaza-Florido A, Amaro-Gahete FJ, Acosta FM, Migueles JH, Molina-Garcia P, Sacha J, Sanchez-Delgado G, and Martinez-Tellez B. Impact of using different levels of threshold-based artefact correction on the quantification of heart rate variability in three independent human cohorts. Journal of clinical medicine 2020; 9:325
21. Connor JT, Martin RD, and Atlas LE. Recurrent neural networks and robust time series prediction. IEEE transactions on neural networks 1994; 5:240–54
22. Zubair M, Mouli GNC, and Shaik RA. Removal of Motion Artifacts from ECG signals by Combination of Recurrent Neural Networks and Deep Neural Networks. *2020 2nd International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*. IEEE. 2020 :1–7
23. Lipponen JA and Tarvainen MP. A robust algorithm for heart rate variability time series artefact correction using novel beat classification. Journal of medical engineering & technology 2019; 43:173–81
24. Azar J, Makhoul A, Couturier R, and Demerjian J. Deep recurrent neural network-based autoencoder for photoplethysmogram artifacts filtering. Computers & Electrical Engineering 2021; 92:107065
25. Baek HJ, Cho CH, Cho J, and Woo JM. Reliability of ultra-short-term analysis as a surrogate of standard 5-min analysis of heart rate variability. Telemedicine and e-Health 2015; 21:404–14

26. Melo HM, Marques JLB, Fialho GL, Wolf P, D'Ávila A, Lin K, and Walz R. Ultra-short heart rate variability reliability for cardiac autonomic tone assessment in mesial temporal lobe epilepsy. Epilepsy Research 2021; 174:106662
27. Canino MC, Dunn-Lewis C, Proessl F, LaGoy AD, Hougland JR, Beck AL, Vaughan GP, Sterczala AJ, Connaboy C, Kraemer WJ, et al. Finding a rhythm: Relating ultra-short-term heart rate variability measures in healthy young adults during rest, exercise, and recovery. Autonomic Neuroscience 2022; 239:102953
28. Medsker LR and Jain L. Recurrent neural networks. Design and Applications 2001; 5:64–7
29. Staudemeyer RC and Morris ER. Understanding LSTM–a tutorial into long short-term memory recurrent neural networks. arXiv preprint arXiv:1909.09586 2019
30. Chung J, Gulcehre C, Cho K, and Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 2014
31. Greff K, Srivastava RK, Koutnık J, Steunebrink BR, and Schmidhuber J. LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems 2016; 28:2222–32
32. Vollmer M, Bläsing D, Reiser J, Nisser M, and Buder A. Simultaneous physiological measurements with five devices at different cognitive and physical loads (version 1.0.1.) Physionet 2022. DOI: 10.13026/zhns-t386
33. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, and Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. circulation 2000; 101:e215–e220
34. Schuster M and Paliwal K. Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing 1997; 45:2673–81. DOI: 10.1109/78.650093
35. Biewald L. Experiment Tracking with Weights and Biases. Software available from wandb.com. 2020. Available from: https://www.wandb.com/
36. Müller M. Dynamic time warping. Information retrieval for music and motion 2007 :69–84
37. Chollet F et al. Keras. https://keras.io. 2015
38. Sheridan DC, Dehart R, Lin A, Sabbaj M, and Baker SD. Heart rate variability analysis: How much artifact can we remove? Psychiatry Investigation 2020; 17:960
39. Kim KK, Kim JS, Lim YG, and Park KS. The effect of missing RR-interval data on heart rate variability analysis in the frequency domain. Physiological measurement 2009; 30:1039
40. Gronwald T, Rogers B, and Hoos O. Fractal correlation properties of heart rate variability: A new biomarker for intensity distribution in endurance exercise and training prescription? Frontiers in Physiology 2020; 11:550572
41. Pan SJ. Transfer learning. Learning 2020; 21:1–2