POLITECNICO DI TORINO

Master of Science in Biomedical Engineering

Master Thesis

Long short-term memory network for heart rate prediction and exercise training load determination using wrist-worn acceleration and heart rate signals in elite runners

Supervisor
Prof. Luca Mesin

Candidate
Carla La Mela

Academic year 2018/2019

Abstract

This study proposes a model to reconstitute heart rate during real running training sessions in elite runners when the optical pulse rate signal is corrupted, commonly due to motion artifacts, not optimal contact skin-sensor, ambient temperature, venous blood flushing and skin tone.

The exploit of wrist-worn optical heart rate monitors in the sport industry market led to a low cost, portable and comfort technology, but the gain in convenience and comfort all in one device came at the expense of accuracy. While for recreational runners the compromise in heart rate accuracy could be acceptable, for elite and sub-elite runners, accuracy can not be compromised because of the relation between heart rate and exercise intensity. The determination of exercise intensity and exercise training load is of paramount importance for elite runners in order to avoid injuries and improve performance.

For this reason, through an accelerometer and an optical sensor, will be tested the feasibility to accurately predict the heart rate through a *Long Short-Term Memory* (LSTM) network.

LSTM is a type of recurrent neural network, able to solve time series task unsolvable by feed-forward networks allowing to make speech recognition, language translation, image recognition and prediction.

The heart rate prediction will be followed by the determination of training load.

Questo studio propone un modello per la ricostruzione del battito cardiaco durante sessioni di allenamento di corsa, qualora il segnale ottenuto tramite sensori ottici sia corrotto, a causa di artefatti da movimento, dal contatto non ottimale pelle-sensore, per la temperatura dell'ambiente, per la presenza di sangue venoso e per il colore della pelle.

Infatti negli ultimi anni, si e' verificato un exploit nel mercato sportivo di sensori ottici indossabili al polso per il monitoraggio del battito cardiaco che ha portato allo sviluppo di una nuova tecnologia economica, portatile e comoda.

La convenienza nell'avere tutto in un unico dispotivo pero' va a spese dell'accuratezza stessa del dispositivo che potrebbe portare eventualmentente ad una non corretta valutazione del battito cardiaco. Se per atleti dilettanti che corrono a scopo ricreativo cio' potrebbe anche risultatre accettabile, per corridori professionali il monitoraggio del battito cardiaco non puo' essere compromesso. Cio' e' dovuto all'esistenza del rapporto tra battito cardiaco e intensita' dell'esercizio.

Conoscere l'intensita' dell'esercizio e del carico di lavoro svolto e' di vitale importanza per corridori professionali, al fine di evitare gravi lesioni al corpo e migliorare la performance fisica.

Per questo motivo, in questo studio verra' testata la fattibilita' di predire accuratamente il battito cardiaco con l'aiuto di un accelerometro e di un sensore ottico attraverso una tipologia di rete neurale chiamata Long Short-Term Memory (LSTM) network, cioe' una rete neurale basata su una memoria a lungo e breve termine.

Tale rete e' in grado di risolvere problemi legati alle serie temporali irrisolvibili con le tradizionali reti neurali, e permette di fare riconoscimenti vocali, traduzioni linguistiche, riconoscimenti in immagini e video, ed appunto predizioni di serie temporali.

La predizione del battito cardiaco sara' seguita dalla determinazione del carico di lavoro svolto.

CONTENTS

1 Introduction 1

1.1 Neural networks 5

1.2 Deep learning 8

1.3 Long Short-Term Memory neural network 10

2 Methods and materials 13

2.1 Study design and subjects 13

2.2 Experimental protocol 14

2.3 Data processing 15

2.4 Linear regression 16

2.5 LSTM model 21

2.5.1 Data manipulation 21

2.5.2 Methods used for data corruption simulation 24

2.5.3 Test set, training set and cross validation 28

2.5.4 Architecture and learning parameters 29

2.6 Metrics for evaluating model performance 30

3 Results and Discussion 32

3.1 Better method for data manipulation 32

3.2 Performance training set and test set 33

3.3 Linear regression prediction and LSTM prediction 35

3.3.1 Error in training load intensity zones 36

3.4 Comparison between HR monitored by OHRM and HR predicted by LSTM 43

3.4.1 Error in training load intensity zones 44

3.5 Literature comparison 49

3.6 Limitations 51

4 Conclusions 55

5 Appendix 57

6 Bibliography 58

1 INTRODUCTION

Heart rate (HR) is a key physiological parameter affected by a number of physiological and behavioral stimuli [1]. Heart rate is linked to exercise intensity by its direct relation with cardiac output [1]. One of the most frequent applications of HR monitoring is to monitor exercise intensity in sports [1] [2] [3] [4].

Monitoring training intensity and training load is of paramount importance for elite runners, on the one hand to improve performance and its determinates, such as, cardiorespiratory fitness, anaerobic threshold, running economy, and on the other hand, to minimize the risk of injury, illness and overtraining [5].

For this reason, typically, training intensity can be displayed in five different HR intensity zones: 50%-60% of maximal HR (HRmax) (zone 1 - very easy intensity), 60%-70% of HRmax (zone 2 - easy intensity), 70%-80% of HRmax (zone 3 - moderate intensity), 80%-90% of HRmax (zone 4 - vigorous intensity), 90%-100% of HRmax (zone 5 - maximal intensity) [6].

ZONE	%HRmax	Intensity
1	50-60%	Very easy
2	60-70%	Easy
3	70-80%	Moderate
4	80-90%	Vigorous
5	90-100%	Maximal

Table 1: Training intensity zones based on maximum heart rate.

According to Edward's summated-heart-rate-zones (SHRZ) model [7][8] , the sum of duration spent in each zone multiplied by a zone factor (1, 2, 3, 4, 5 respectively for each zone) can determine the training load. Edward's training load [7] [8]can be calculated as follows:

$$TL = duration\ zone1 \times 1 + duration\ zone2 \times 2 + duration\ zone3 \times 3 + \\ + duration\ zone4 \times 4 + duration\ zone5 \times 5 \qquad \textit{Equation 1}$$

Where each duration refers to the duration in minutes.

Endurance runners need to alternate long and low intensity trainings, so called volume trainings, which enhance general central capacity, to shorter and higher intensity trainings, also referred to intensity trainings or quality trainings, aiming at improving running economy, and racing speed [9]. This is why runners and coaches find rather useful using HR zones. When considering the distribution time spent in each zone 5 different training types can be derived: a. prolonged high-volume training with low intensity (HVT); b. low volume high intensity interval training (HIIT); c. combined training, also called polarized training (POL); d. threshold training (THR); and finally e. equal distribution "uniform" training (UNI) [6]. The training type information is very valuable for training periodization.

The sport industry started to exploit HR monitoring in the 1980's with the development of wireless, wearable HR monitors consisting of an ECG-based chest strap sensor and a wristwatch radio receiver [1]. However this 30 years old technology, has now been disrupted by less obtrusive strapless HR monitors, based on photo-plethysmography (PPG) and implemented directly on the wristwatches[1][3][10][11].

In the last 5-6 years there has been an exponential increase of wrist-worn optical HR monitors in the sport industry market due to the low cost, portable and comfort technology [1][3][10][11].

The gain in convenience and comfort all in one device came at the expense of accuracy[12]. Indeed ECG-based wireless chest straps still show the closest agreement to conventional lead ECG gold standard. As we directly experienced in the field, if for recreational runners the compromise in HR accuracy could be acceptable, for elite and some sub-elite runners, accuracy cannot be compromised.

Photoplethysmography illuminates the skin via light-emitting diode measuring the intensity of the reflected light to the photo-detector. The intensity of the reflected light follows the blood volume changes in the arterial vessels caused by the pressure pulse of the cardiac cycle permitting to obtain heart rate [11][3].

The output of the PPG sensor is often affected by a low signal-to-noise ratio (SNR), due to several reasons: first of all because of motion artifacts especially in running where the number and the entity of subject-sensor movements is really high [13]. Furthermore, the sensor location is the

medial part of the dorsal wrist that, in runners, is often concave, squared and bony because they are really skinny and its anatomical shape not always fit well with the mechanical design of the device, causing a non optimal contact skin-sensor [1] . In addition, ambient temperature can affect PPG accuracy. When the environment is cold, blood vessels are more constricted, in order to reduce heat dissipation and preserve the optimal core temperature, resulting in a reduction in sub-cutaneous blood perfusion [11]**.** Venous blood flushing is a less known source of PPG artefact, but it is believed to play an important role too. This consists in the noise produced by venous blood, which per se is not pulsatile, that when the arm moves rapidly can flush creating an artefact [11]. Finally, skin tone could also influence PPG-based HR accuracy. Different skin tones absorb light differently, for example, darker skins absorb more green light reflecting a small amount of light to the detector [14]. Thus, "it is generally accepted that PPG-based HR monitoring suffers from inherent drawbacks" [12].

As anticipated above, due to the lower accuracy of PPG sensor during running trainings, sub-elite and elite athletes (e.g. runners) who need accurate training intensity and training load determination, are reluctant to adopt the wrist-worn HR monitors, preferring still traditional chest-strap HR monitors [1].

Clearly if the accuracy in determining training intensity and training load could be improved those runners may seriously consider adopting the more convenient wrist-based technology.

Thus the purpose of this study was to test whether a wrist-worn accelerometry-derived feature, such as activity counts (ACN), feed into a Long Short-Term Memory (LSTM) networks, could accurately reconstitute heart rate (HR) and consequently training intensity and training load, during real running training sessions of elite runners, when the optical pulse rate signal is corrupted because of a small signal to noise ratio, commonly due to motion artifacts and the list of artifacts mentioned above.

LSTM is a type of recurrent neural network, able to solve time series task unsolvable by feed-forward networks allowing to make speech recognition, language translation, image recognition and prediction[15]. LSTM network have memory blocks connected through layers allowing the network to learn basing on the history of the time sequence data.

Interestingly, several earlier attempts have been made to try to correct HR measurements from motion artifacts through signal processing techniques. For example using adaptive noise cancellation[16][17], accelerometer used to estimate frequency and the noise produced by motion removed from HR through a noise canceler[3][18] and a combination of two algorithms for removing motion artifacts and a spectral peak tracking [19].

In the last years also machine learning approaches have been already used for HR correction.

Jindal [20], used a wrist-worn watch with a PPG and a tri-axial accelerometer, built a deep learning classification model for determining heart rate, where motion data were used only to filter the signal before input the heart rate features in the model. The difference with our study is that acceleration is not directly used to estimate the HR

R. McConville *et al.*[21] built a regression model using acceleration data together with heart rate data for heart rate prediction in patients with aortic and mitral valve disease. This was a different situation compared to our study, patients were not doing training activity and HR was not corrupted. The task of their study was to reconstruct heart rate during daily life with an accelerometer instead of using a PPG sensor, in order to gain energy from wearable devices, because PPG foe working uses more energy than accelerometer.

Y. Ming and J. Jun [22] built a classification feedforward neural network using acceleration data and heart rate data for heart rate prediction in 90 minutes signal from a single healthy male. Heart rate was recorded during daily life through a portable HR monitor with electrodes and accelerometer data were recorded through a tri-axial accelerometer. In this case, HR prediction was not performed during running activity but daily activity and the model was subject-based.

In this study we propose a novel method for heart rate prediction using accelerometer feature during running exercise training in a LSTM network.

1.1 NEURAL NETWORKS

In 1956, a new field called *Artificial intelligence* spread [23], with the purpose to develop machines that could simulate aspects of human intelligence.

In 1958, the first and the simplest neural network was implemented with the name *Perceptron* by Frank Rosenblatt [24].

The Perceptron model is a binary classifier consisting of an input vector $\mathbf{x} \in \mathbb{R}^n$ with an associated weight vector $\mathbf{w} \in \mathbb{R}^n$. Its binary output function $\varphi(\mathbf{w}, \theta) : \mathbb{R}^n + 1 \rightarrow \{-1, 1\}$ with a treshold θ, outputs 1 if $\mathbf{w} \bullet \mathbf{x} \geq 0$ and outputs −1 otherwise [24] [25] [26].

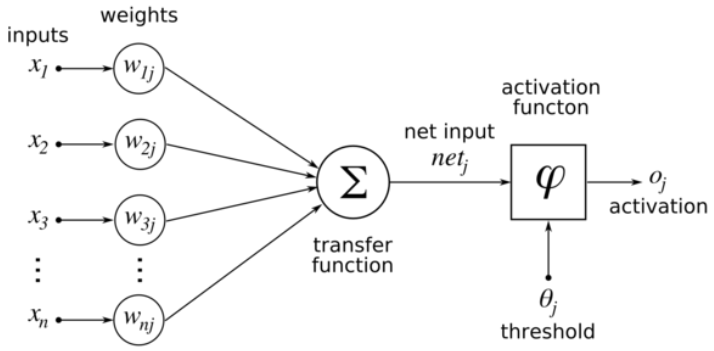


Figure 1: Single layer perceptron, first model of NN [27]*: the neuron receives input through weighted connection, the weighted sum is compared to a threshold and transformed in output by an activation function in one of the two classis.*

Neural networks (called also Artificial Neural network ''ANN'' or ''NN'') are systems, taking inspiration from the human brain, where the main unit is the neuron. Connections between neurons, like synapses in a human brain, can transmit an information from a neuron to another allowing the network to learn through a learning algorithm [24]. Learning algorithm is the procedure used to perform the learning process, it can be supervised or unsupervised. It is supervised if the NN is provided with a dataset consisting of input vectors and target vectors [28], it is unsupervised if the NN is provided only of input vectors [29].

The limitation of *Perceptron* was the simplicity: a single neuron classifying linearly the input in one of two classis. To face the problem of data not linearly separable, in 1986 *Multilayer Perceptron*

was born, with a more complex architecture composed by several neurons connected among them and organized in layers, overcoming the limitations of perceptron. The name of this new type of architecture is *feedforward neural network* [30]. It is composed by an input layer, one or more hidden layer fully connected with the previous layer composed by different numbers of neurons able to learn the information coming from the input layer and finally an output layer with a number of neuron equal to the number of classes [31].

This learning algorithm is called *backpropagation*, is a *gradient based* optimization method useful for finding the optimal set of weights [31]. The name *gradient based optimization* comes from the need to find an optimum where the error between output and the target output is minimized [31]. When the optimum is reached, the learning algorithm converges [31]. To reach the optimum, is necessary a parameter called *learning rate.* Is a parameter with values from 0 to 1, an high learning rate will takes a faster convergence but with the problem of jumping over minima, a too low learning rate takes too long to converge with the problem of reaching non-optimal minimum [31]. The gradient is calculated over the *cost function* (or *loss function*) that is the error between output and the target output [31]. The *cost function* could be, for example, the mean-squared-error (MSE).

The learning algorithm has two phases [31]:

1. Propagation phase: the input information is propagated through layers until it reaches the output layer, error or *cost function* is calculated between the output and target output.
2. Update phase: the gradient of the cost function is computed in order to minimize the error, multiplying this gradient by the learning rate, the weights in the net are updated.

$$w^{n+1} = w^n - \alpha \, \nabla J(w) \qquad \textit{Equation 2}$$

Where w^{n+1} is the updated weight, w^n is the weight to be updated, α is the learning rate, $\nabla J(w)$ is the gradient of the cost function [31].

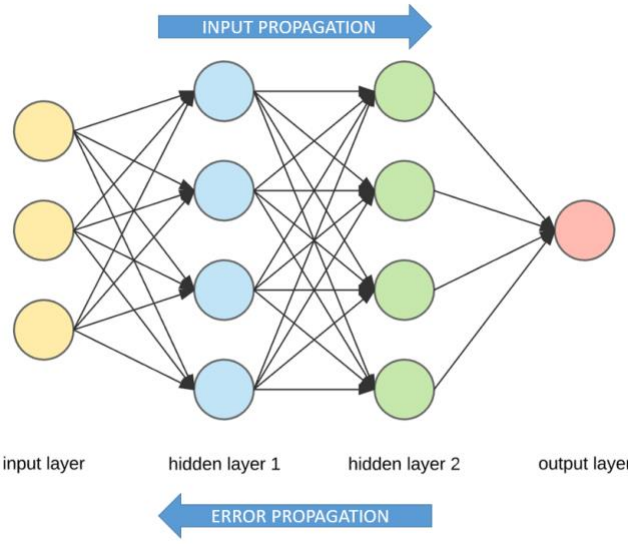


Figure 2: Evolution of single layer perceptron: feedforward neural network with backpropagation optimization.

In machine learning, a challenge to obtain a good model is avoiding overfitting. Overfitting occurs when the network is not able to generalize the data, memorizing the training pattern [32]. Overfitting may be caused by a too complex architecture, the training set size and the optimization method used.

To avoid overfitting, some form of regularization are used:

- Early stopping [33]: from training set a subset is obtained, the validation set. This set help the network to converge as long as until the optimum is reached for the validation error. In this way, the network stops training at the point in which the validation error is minimized and weight parameters are stored.

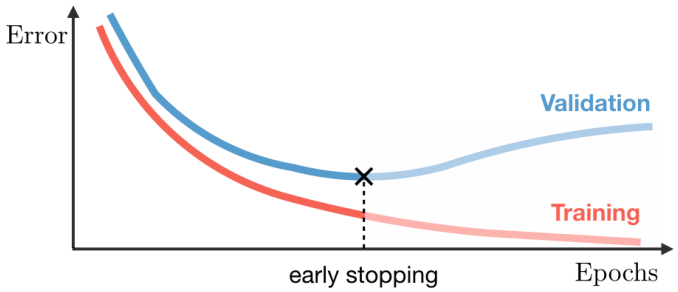


Figure 3: Typical training and validation error during a network training.

Early stopping occurs when validation error optimum is reached.

- Dropout layer [34]: it removes randomly, with a certain probability, unit interconnections in the network for each training iteration.

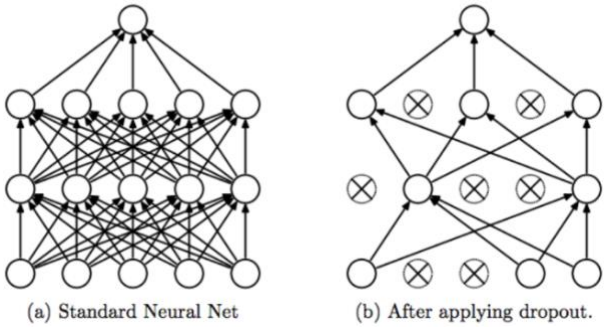


Figure 4: Influence of dropout layer: it removes randomly, with a certain probability, interconnections in the network.

1.2 DEEP LEARNING

In the last two decades, a subfield of machine learning based on neural networks, started to become popular, this new approach is *deep learning*. In the supervised learning field, a deep learning algorithm is able to learn more complex input-output relationship. In particular, through supervised learning is possible to solve both classification and regression problem. For a classification model, the task is to predict a discrete output (called also *class*, *category* or *label*) for a given input. For a regression model, the task is to predict a continuous output variable for a given input.

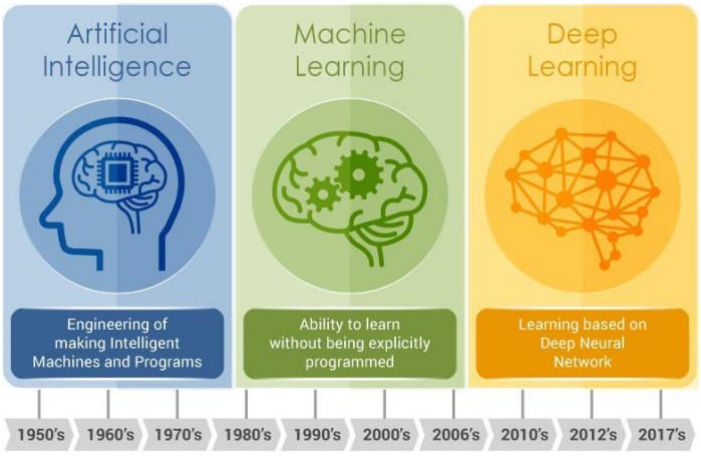


Figure 5: Evolution from artificial intelligence to deep learning

Two types of layers distinguish deep learning from the standard fully connected layer of the old approach: convolutional and recurrent layer. The network based on convolutional layers (called *Convolutional Neural Network* or *CNN*) are mainly popular for image processing, that one based on recurrent layers (called *Recurrent Neural Network* or *RNN*) are used for modelling time series data and sequence data to make speech recognition, language translation and time series prediction.

In particular, we will focus on the last one, the Recurrent Neural Network. Compared to the NN, RNN have the concept of memory: adding a loop in a feed-forward neural network, the network get this behavior of sequential of memory (*Figure 6).* The information passes from one step to the next one in a ordered way. The previous step is called hidden state that acts as a memory holding the information of the previous step.

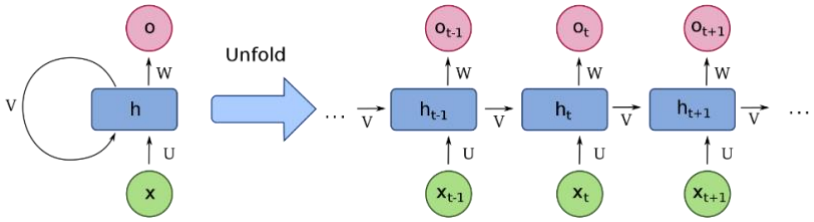


Figure 6: Recurrent neural network with the fold and unfolded structure. The fold structure is similar to a feed-forward neural network adding a cycle allowing the network get this behavior of sequential of memory. The information passes in an ordered way through time how is possible to see in the unfolded structure.

The issue affecting the RNN is the short-term-memory [35]. Learning long dependencies for RNN is difficult because of the vanishing gradient due to the backpropagation algorithm. The updating of weights depends on the gradient, the bigger the gradient the bigger the update and vice versa.

In this way, if the adjustment to the layer before is small because of a small gradient, the following layer will be updated even smaller with any learning results. This is the problem of the vanishing gradient, in which it shrinks exponentially during the backpropagation and RNN is not able to learn long-term dependencies [35].

To face this problem, Long Short-Term Memory (LSTM) neural network were introduced by Hochreiter and Schmidhuber [35] .

1.3 Long Short-Term Memory neural network

LSTM is a type of RNN but able to learn long and short term dependencies [36]. A LSTM layer is formed by blocks (or cells) with internal mechanisms called *gates* able to learn time dependencies [35][36]. The gates can learn the information deciding if keep it or forget it during the network training.

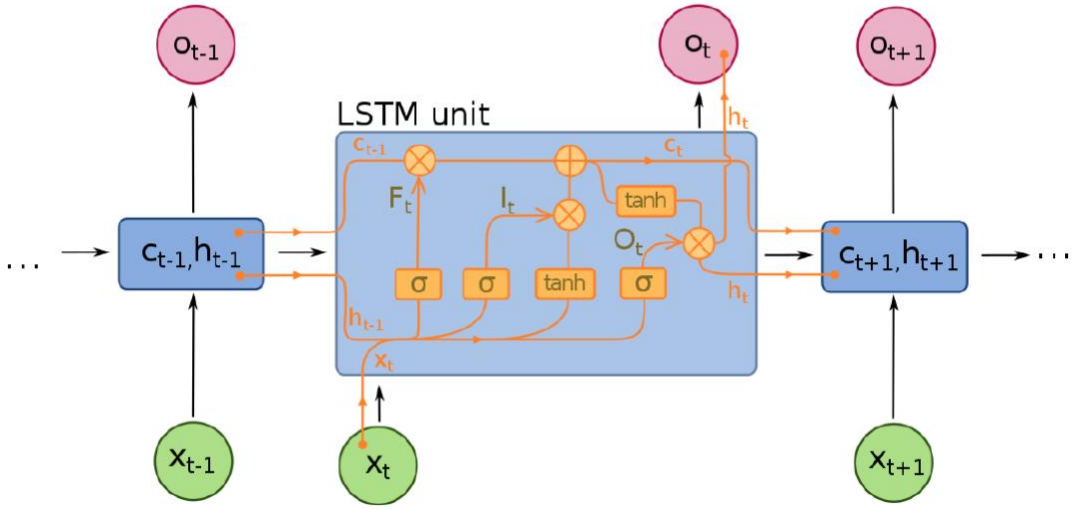


Figure 7: Structure of a LSTM layer formed by blocks called also cells memory with gates able to learn.

Inside the LSTM cells there are mainly three gates: a forget gate, an input gate and an output gate.

These gates use two types of activation functions, the sigmoid and the hyperbolic tangent, defined respectively as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$ *Equation 3*

$$tanh(z) = \frac{sinh(z)}{cosh(z)} = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$ *Equation 4*

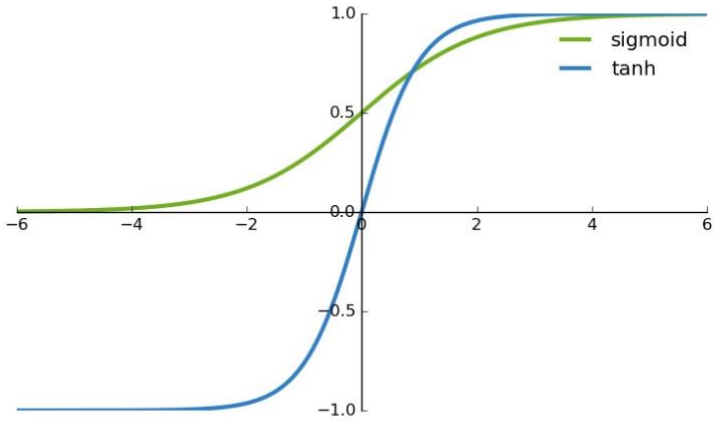


Figure 8: Sigmoid and hyperbolic tangent activation functions

For each gate, three different learnable weigths are considered, W_j, R_j, b_j, respectively the input weights, recurrent weights and bias. Where $j = f, i, o$ index of the three gates.

In the first step, current cell input x_t and the information of the prevous hidden state h_{t-1} pass through the sigmoid function that, giving an output in [0,1], it decides if keep the information or forget. If the value is close to 0 the information will be forgot, instead, for a vaue close to 1 the information will be kept. This function is called forget gate f_t.

$$f_t = \sigma\left(W_f x_t + R_f h_{t-1} + b_f\right)$$ *Equation 5*

In the next step, the hidden state h_{t-1} and the current cell input x_t pass through another sigmoid function that, giving an output in [0,1], it decides which value will be used for the update. This function is called input gate i_t.

$$i_t = \sigma\left(W_i x_t + R_i h_{t-1} + b_i\right) \qquad \textit{Equation 6}$$

The same hidden state h_{t-1} and the current cell input x_t pass also through the hyperbolic tangent function that, giving an output in [-1,1], it regularize the network.

$$\tilde{C}_t = tanh\left(W_c x_t + R_c h_{t-1} + b_c\right) \qquad \textit{Equation 7}$$

The output of the sigmoid function i_t is multiplyed by the output of hyperbolic tangent (tanh) $\tilde{C}_t$ function and the cell state C_{t-1} is multiplied by the forget vector f_t. These two multiplication result are summed creating the new cell state C_t.

$$C_t = i_t * \tilde{C}_t + C_{t-1} * f_t \qquad \textit{Equation 8}$$

The hidden state h_{t-1} and the current cell input x_t pass through another sigmoid function . This function is called output gate o_t.

$$o_t = \sigma\left(W_o x_t + R_o h_{t-1} + b_o\right) \qquad \textit{Equation 9}$$

The new cell state C_t passes through the *tanh* function and is multiplied by the output gate o_t giving the hidden state of the current cell h_t that will be carried over the next cell.

$$h_t = o_t * tanh(C_t) \qquad \textit{Equation 10}$$

Through this memory cells mechanism, LSTM is able to work well with long time sequence series both for classification and regression tasks. For classification predicting or a sequence (sequence-to-sequence classification) or a label (sequence-to-label classification). For regression, instead, involves a prediction of a value (sequence-to-one regression) or a prediction of a sequence (sequence-to-sequence regression).

2 METHODS AND MATERIALS

2.1 STUDY DESIGN AND SUBJECTS

The study protocol was approved by the *Internal committee biomedical experiments* (ICBE) board of Philips Research Eindhoven and tests were conducted at the Philips Research Laboratories, in Eindhoven (NL).

Five elite runners of white ethnicity, two female and three male, have been recruited voluntarily from the local sport running center (

Table 2).

All subjects have signed the informed consent form before starting any type of research activity. Then, they filled an ACSM Health/Fitness Facility Pre-participation screening questionnaire and an overtraining questionnaire was filled in (DALDA).

After no diseases and no suspected overtraining were identified by the questionnaire, each subject were asked to wear a wrist-worn PPG-based HR monitor, here referred to as the optical heart rate monitor (OHRM, Philips Research, The Netherlands). The subjects were also asked to wear a wearable metabolic system (K5, Cosmed, Italy), a device used for measuring metabolic parameters (such as VO2, VCO2, ventilation, HR, energy expenditure etc…) in order to measure the maximal oxygen uptake (VO2max) and the anaerobic threshold of each athlete.

The OHRM contains a tri-axial accelerometer (sensitivity=256 lsb/g, range=±8g, sampling frequency of 128 Hz) and a PPG sensor (sampling frequency of 128 Hz, light-emitting diode at 530nm wavelength) .

Subjects were asked to fill two diaries and one questionnaire in order to get more information during these 15 days (see *5*)*:* a morning diary for having info about daily life's athlete (filled every morning), a training diary to get info about training activity (filled every training session) and a recovery questionnaire to get info about their recovery status (filled at the end of each week).

Table 2: Mean (±SD) characteristics of the subjects

Age (yr)	25±7.31
Height (cm)	178.6±13.58
Weight (kg)	60.62±9.67
HR rest (bpm)	57.6±9.04
HR max (bpm)	186.±13.9
Anaerobic threshold (bpm)	178±15.28
VO_{2max} (ml/kg*min)	74±5.79

2.2 EXPERIMENTAL PROTOCOL

A VO2max test was conducted on a treadmill (*Excite run 600, Technogym*, Italy). VO2max test consists of an incremental intensity exercise test that leads the athlete to reach the maximum oxygen consumption. Speed, duration and treadmill inclination for warming up was chosen by the athlete until he/she felt ready to start the test. The warmup could not last more than 10 minutes.

The incremental test consisted of a fixed incline at 3%, with a starting speed of 8 km/h for women and 10 km/h for men. Every minute the speed was increased of 1km/h until the subject decided to stop because of achieving exhaustion. This was checked off-line by assign the measure of heart rate maximum (HRmax), being greater than 90% of the aged estimated HRmax according to Tanaka et al. formula [37]:

$$HRmax = 206 - 0.7 * age \qquad \textit{Equation 11}$$

and checking the respiratory exchange ratio (amount of carbon dioxide over the amount of oxygen) exceeding > 1.15 [38].

Speed range, during the VO2max test, was from 8km/m until 21km/h with a final mean speed reached of 18.8±2.18 km/h. The test was followed by a cooling down period with speed, duration and inclination chosen by the athlete.

Participants were asked to wear the OHRM for 15 days (24/7), in order to get their free-living training data and also an ECG-based chest strap (only during trainings) in order to get heart rate reference.

After these 15 days were passed, another VO2max test was conducted.

2.3 DATA PROCESSING

Data were transferred via USB into a personal computer and with a software developed by Philips, row data from PPG sensor and tri-axial accelerometer were processed in order to produce accelerometry and PPG derived features such as heart rate (HR) and activity counts (ACN).

Activity counts is obtained from the sum of the integral of the absolute acceleration for each axis over 1s interval [39].

$$ACN = \int_{t_0}^{T+t_0} |a_x|\ dt + \int_{t_0}^{T+t_0} |a_y|\ dt + \int_{t_0}^{T+t_0} |a_z|\ dt \qquad \textit{Equation 12}$$

Where a_x, a_y, a_z are the acceleration on the three axis, $[t_0, T + t_0]$ is 1s interval.

These features were processed and analyzed using Matlab R2019a (Mathworks, Cambridge, MA, USA) using a GPU-based workstation.

From 24h data for fifteen days of each of the five subjects, with the help of subject's training diary and through an accelerometer feature (activity type) containing index of the type of activity

recognized (possible activities were walking ,running, cycling and other) and also by visual inspection, 45 running sessions were collected with a total duration of 37 hours and 48 minutes (50± 30 minutes), containing 10 VO2max test (2 for each subject), 2 races, 16 interval trainings, 18 long run. The other days were of resting or other sport activities (cycling, yoga, core and strength exercise, etc...).

2.4 LINEAR REGRESSION

At first, a linear regression was performed between activity counts (ACN) and heartrate reference (HR) during running training sessions, to model the relationship between these two variables.

Given a dataset, composed by two variables (x, y), a simple linear regression model assumes that, between these two, exists a linear relationship [40].

The simple linear regression line formula is:

$$y = mx + q \qquad \textit{Equation 13}$$

Where y is the dependent variable, x is the independent variable, m and q are called regression coefficients, respectively the slope and intercept of the regression line.

The correlation between two random variables is expressed by Pearson's correlation coefficient, is a measure of linear correlation of two variables x, y [41].

$$\rho_{x,y} = \frac{cov(x, y)}{\sigma_x \sigma_y} \qquad \textit{Equation 14}$$

Where $cov(x, y)$ is the covariance between the two variable x and y, and σ_x, σ_y are the standard deviation of the two variables. Values of Pearson's correlation coefficient vary from -1 to 1.

Positive values of $\rho_{x,y}$ reflect two variables directly proportional, negative values refer to a relation indirectly proportional between the two variables x, y. Furthermore, values close to 1 and -1 indicate a strong correlation, values close to 0 indicates the absence of correlation between the two.

Pearson's correlation coefficient reflects the strength of a linear relationship between two variables, but doesn't give information about non-linearity relationship between the independent variable x and the dependent variable y [42].

Computing correlation coefficient between the two variables, it results to be: $\rho_{x,y} = 0.72$.

This results show that ACN and HR are strongly related.

The scatterplot in *Figure 9* shows HR data (along y-axis) and ACN data (along x-axis), of the 45 running sessions of the five subjects and in red is showed the fitted regression line.

The distance from the points of the dataset to the regression line is called error ε. With the error, the equation become:

$$y = mx + q + \varepsilon \qquad \textit{Equation 15}$$

If the error is zero, no distance separates dataset points to the regression line, as a result that all the points are well fitted into the regression line. If (x, y) fits the linear regression line, the relation between them is perfectly linear.

Looking into the scatterplot, ACN and HR do not show a mere linear relation: is possible to see two big clouds with many dispersed points not well fitted into the linear regression line, as a result that exists a non-linear relationship between ACN and HR.

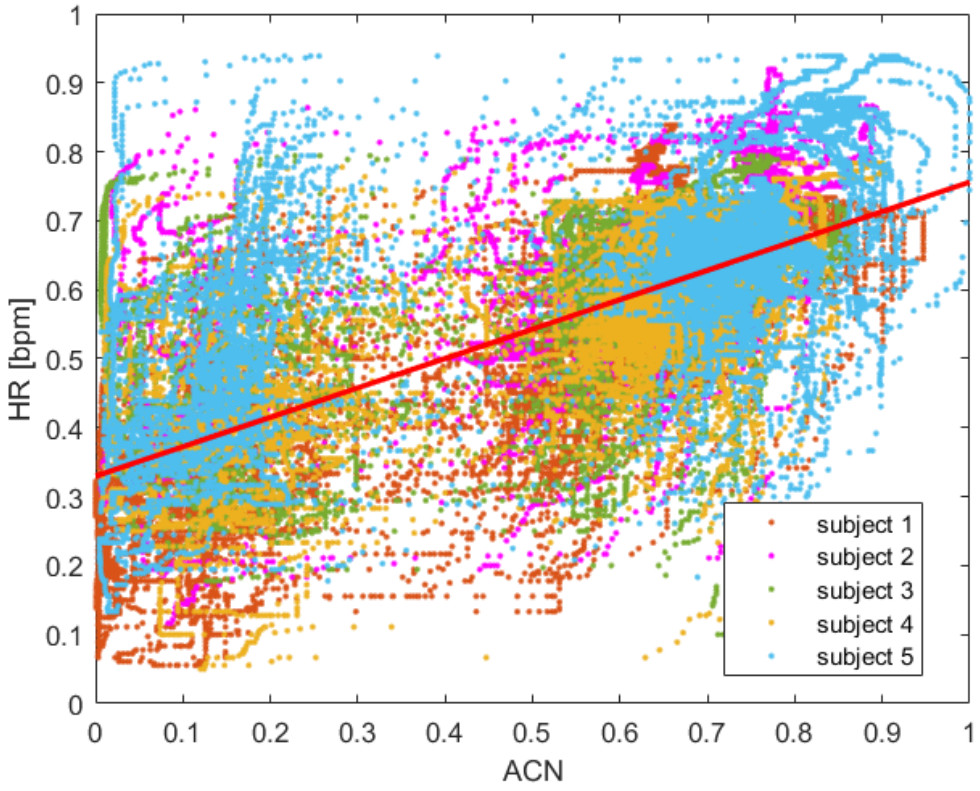


Figure 9: Linear regression scatter plot between ACN and HR by subject-color with normalized values. Correlation coefficient ($\rho_{x,y}$ = 0.72), shows that ACN and HR are strongly related, the scatterplot doesn't show a linear relation between them. Is possible to see two big clouds surrounded by many dispersed points. These dispersed points, not well fitted into the red linear regression line, are due to the non-linear relationship between ACN and HR.

The relationship can be affected by different factors: subject-dependent factors (such as age, fitness level, hydration, internal temperature, etc…), different kind of running trainings that can change the linearity of the relation, the kinetics between acceleration and heart rate, especially talking about running, are different. If we consider the beginning of a run, the acceleration has an instant rise while the heart rate rises with a delay compared to the acceleration response (this can be explained by the cloud of points on the right side of *Figure 9* under the regression line). The same happens at the end of a run, while the acceleration has an instant decrease, heart rate is still high decaying with a delay (explained by the cloud of points on the left side of *Figure 9* above the regression line).

The linear relation can be also influenced by the cardiovascular drift, it is the time dependent change of the cardiovascular response [43] where heart rate rises even if the workload is not increasing (cloud of points on the upper part of *Figure 9* above the regression line is explained by this behavior). It can be influenced by subject-dependent factors as well (internal temperature, hydration and the amount of muscle tissue involved during exercises) but also by ambient temperature [43].

Figure 10 shows a cardiovascular drift event and the different kinetics between ACN and HR during a training session. In orange line the ACN and in blue line the HR. The beginning of the training is characterized by an instant rise in ACN and a delayed rise in HR (n.1). The opposite ad the end of the running while ACN decreases instantaneously and HR has a delayed decrease (n.2). Furthermore, in the central part of the session, while ACN is almost constant, HR rises without an increase in workload, this is a sign of cardiovascular drift (n.3).

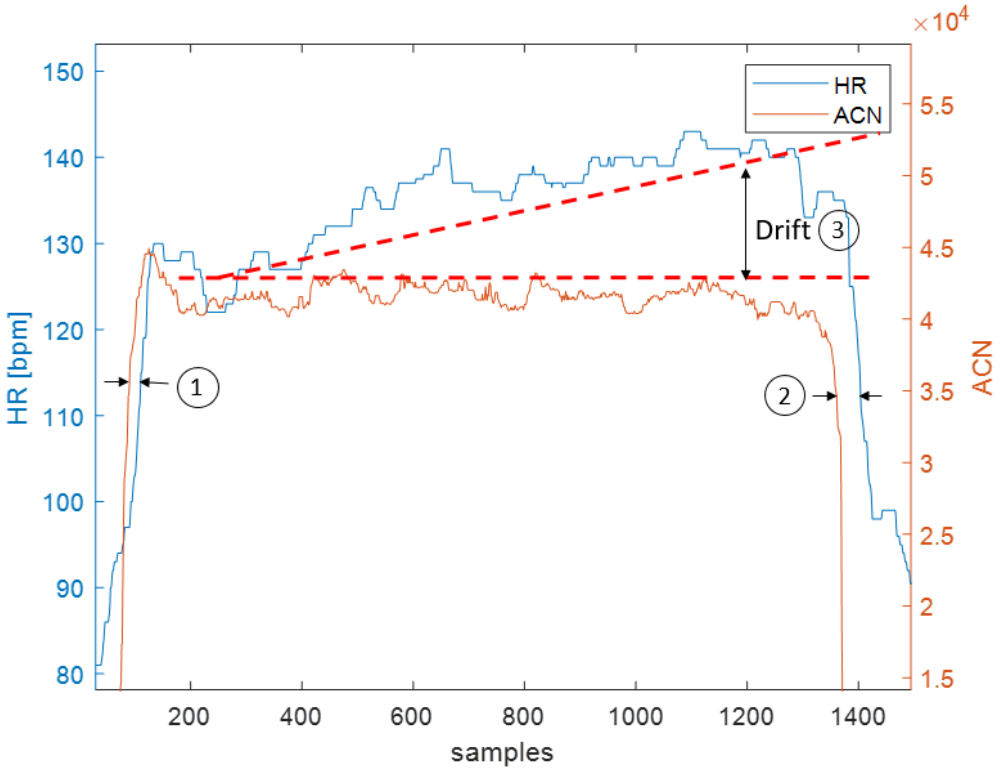


Figure 10: Cardiovascular drift event and the different kinetics between ACN and HR during a training session. In orange line the ACN and in blue line the HR. The beginning of the training is characterized by an instant rise in ACN and a delayed rise in HR (n.1). The opposite happens at the end of the running while ACN decrease instantaneously and HR has a delayed decrease (n.2). Furthermore, in the central part of the session, while ACN is almost constant, HR rises without an increase in workload, sign of cardiovascular drift (n.3).

Pearson correlation coefficient can be affected by some factors: amount of variability in the data, differences in the shape of the two distributions, lack of linearity, the presence of one or more outliers, characteristics of the samples and measurements errors [42]. For example two variables can have a weak value of correlation, but their relation can be statistically significant and vice versa.

For this reason, to test the significance of the correlation coefficient, a null hypothesis test has been performed.

Null hypothesis test consist of considering $\rho_{x,y}$ equal to zero, this means considering no relationship between (x, y) variables:

$$H_0 : \rho_{x,y} = 0 \qquad \textit{Equation 16}$$

In the context of null hypothesis $\mathrm{p-value}$ (probability value) is used to quantify the statistical significance between to variables [44], considering a significance level of 95% ($\alpha = 0.05$), $\mathrm{p-value}$ results to be: $p < 0.05$.

It is possible to conclude that the null hypothesis can be rejected, ACN and HR are statistically significantly moderately related.

Another important parameter used to understand how the independent variable explains the variability of the dependent variable is the coefficient of determination. It can be calculated as the square of Pearson correlation coefficient [45]:

$$R^2 = (\rho_{x,y})^2 \qquad \textit{Equation 17}$$

Range of determination coefficient are from 0 to 1, in our case it is equal to: $R^2 = 0.52$

This means that the 52% of the variability of dependent variable is depends on the independent variable, the remaining 48% of the variability of y depends on other unknown factors.

Computing the formula of the regression line with the given dataset $[HR(t), ACN(t)]$, where heart rate reference is the dependent variable and activity counts is the independent variable, the parameters of the linear model are estimated: $HR(t) = 0.43\, ACN(t) + 0.33$.

Linear regression can be used to fit a predictive model. After have developed a predictive linear model, this can be used to make the prediction of the independent variable.

This equation will be used to predict linearly heart rate values having activity counts varying during time for comparing the improvements of an LSTM model.

2.5 LSTM MODEL

2.5.1 Data manipulation

Only 30 over 45 running sessions were corrupted because of the low accuracy of the OHRM.

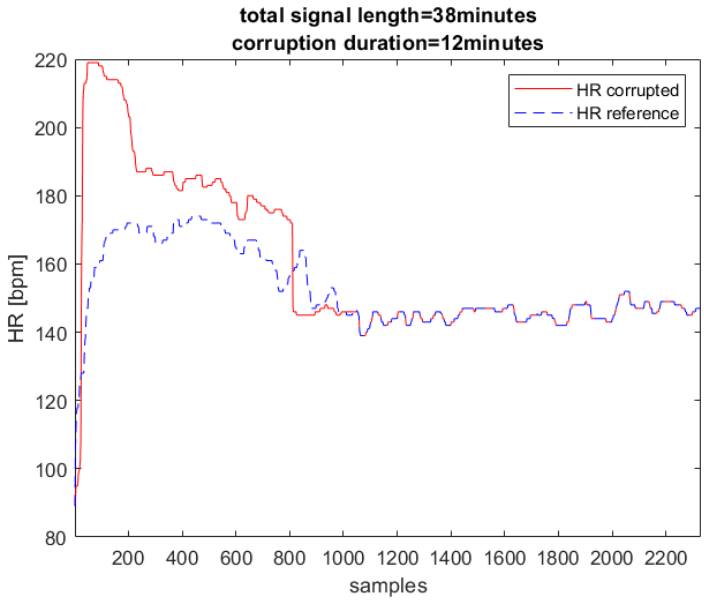
Data augmentation was applied to our dataset, composed by 45 training sessions, to achieve a more generalized model avoiding overfitting problems due to the training set size already mentioned in *section 1.1*.

Data augmentation is a technique becoming always more common with the spread of deep learning, to avoid the limited amount of data often causing overfitting [46] [47]. The approach consist in or generating data from scratch or perturbing existing data creating new ones [46]. For example Oksuz and Ruijsink [48], to build a convolutional neural network for image motion artefact detection, performed a data augmentation corrupting images with realistic motion artefact. The same was performed by Hoffmann [49] corrupting images with different percentage of random noise to develop a classification prediction model.

In this study, data augmentation was performed simulating a certain part, of the 45 heart rate reference signals, corrupted in random samples. Corruption duration was chosen from 5 to 40

minutes, based on the real corruption of the OHRM signals (19.59 ± 19.82 minutes according to the OHRM HR estimation algorithm).

A)



B)

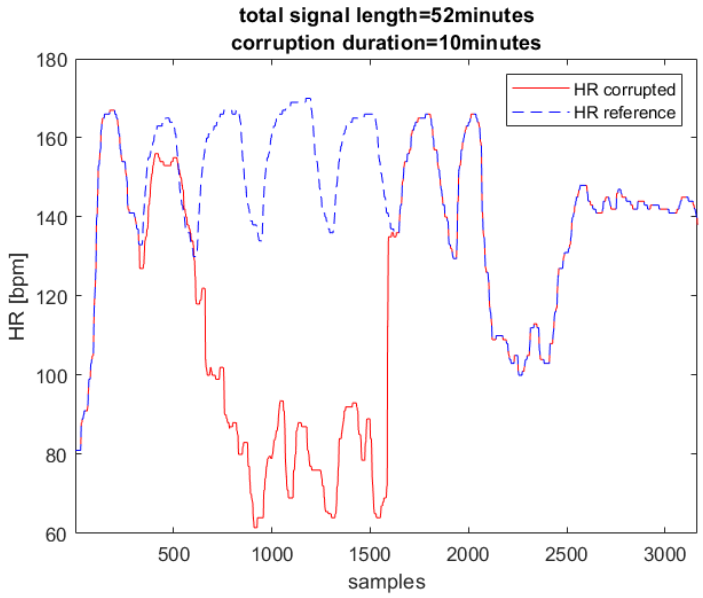


Figure 11: Examples of sessions out of 45 training sessions of the dataset: in red line a real case of corrupted HR recorded with the OHRM, in blue line HR reference recorded with the ECG chest-strap. A) Long run where is clearly visible how HR coming from OHRM is not reliable, reaching a peak of 220 bpm, value extremely too high. The duration of corruption equal to 12 minutes over 38 minutes of total training duration. B) Interval training. In this case, starting from the second interval, HR starts to decrease in accuracy reaching values of 60 bpm, a too low value to reach during an interval training, almost close to the subject heart rate rest. Duration of corruption equal to 10 minutes over 52 minutes of total training duration.

All the 45 training sessions have different duration, for this reason, corruption performed was dependent on the duration of each HR signal:

- For signal duration> 42 minutes, corruption was performed five times, for 40, 30, 20, 10 ,5 minutes, allowing to obtain 5 heart rate signals randomly corrupted from one.
- For 32 minutes $<$ signal duration $\leq$ 42 minutes, corruption was performed four times, for 30, 20, 10, 5 minutes, allowing to obtain 4 heart rate signals randomly corrupted from one.
- For 22 minutes $<$ signal duration $\leq$ 32 minutes, corruption was performed three times, for 20, 10, 5 minutes, allowing to obtain 3 heart rate signals randomly corrupted from one.
- For 12 minutes $<$ signal duration $\leq$ 22 minutes, corruption was performed two times, for 10, 5 minutes, allowing to obtain 2 heart rate signals randomly corrupted from one.
- For signal duration $\leq$ 12 minutes, corruption was performed only once for 5 minutes, obtaining one signal.

Form 45 training session, the dataset was augmented obtaining finally 176 running sessions.

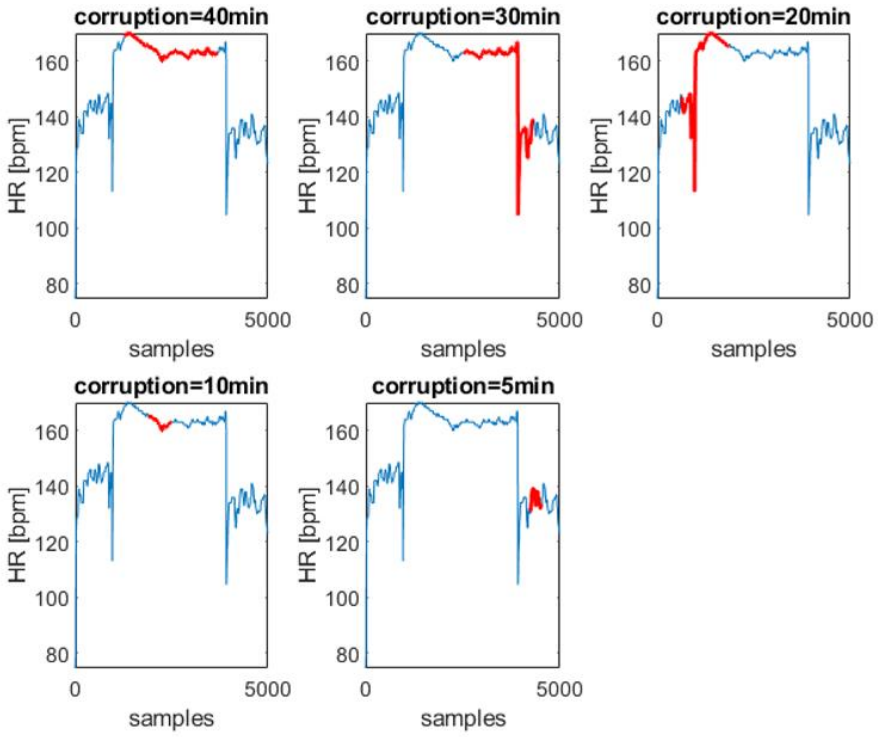


Figure 12: Example of random samples corruption in HR data reference. From one training session, five training sessions corrupted differently are obtained.

2.5.2 Methods used for data corruption simulation

In the previous step was analyzed the way to obtain more training sessions from a smaller dataset, and to be more precise, where the corruption was (randomly samples selected) and for how long.

The corruption was performed testing three methods in the same corrupted samples.

Let's consider ACN and HR of a training session and random corrupted samples from 2000 to 3000 (see *Figure 13*).

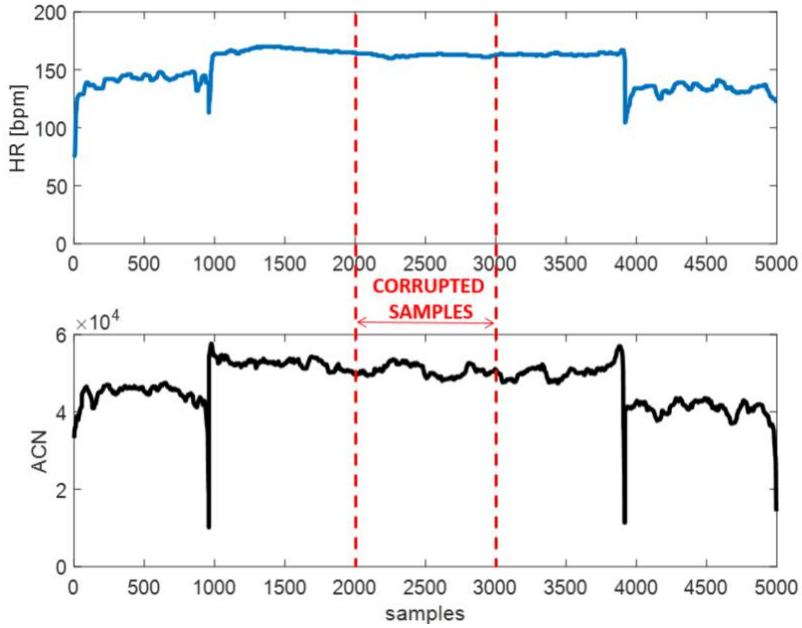


Figure 13: ACN (black signal) and HR (blue signal) are considered from the same training session whose random samples considered are from 2000 to 3000.

First of all, pre-processing was performed over ACN, that will be the same for all the three methods. In order to have ACN signal only in the zone of interest (i.e. where the HR is simulated corrupted) zero padding was performed in non-corrupted samples, to emphasize the period of corrupted parts in heart rate signal.

Considering the previous example, ACN values will be different to zero only in corrupted samples (from 2000 to 3000) and zero in the remaining samples (from 1 to 1999 and from 3001 to 5000).

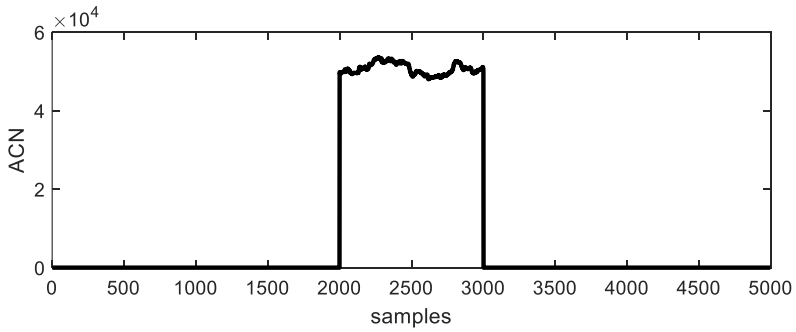


Figure 14: Pre-processing was performed over ACN, that will be the same for all the three methods. In order to have ACN signal only in the zone of interest (i.e. where the HR is simulated corrupted) zero padding was performed in non-corrupted samples, to emphasize the period of corrupted parts in heart rate signal. Considering the previous example, ACN values will be different to zero only in corrupted samples (from 2000 to 3000) and zero in the remaining samples (from 1 to 1999 and from 3001 to 5000).

1) HR ZERO method

The first method consist of performing the corruption on HR with a zero padding in corrupted samples:

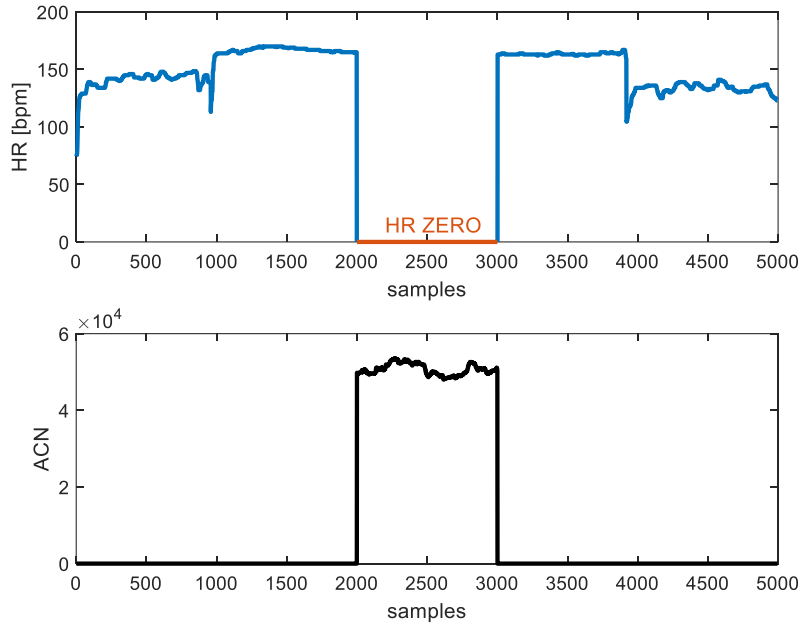


Figure 15: HR ZERO method: On the top figure a zero padding (orange line) was performed over HR (blue signal) in corrupted samples. On the bottom figure, ACN is always the same as was previously explained.

2) HR LAST method

Corruption was performed in heart rate data inserting last values before the corruption starts.

Having a frequency sampling of 1Hz, the last value will be 1s before the corruption starts.

For example, considering the same range of corrupted samples (from 2000 to 3000), HR in this interval will have the value at sample 1999, 1s before the starting interval.

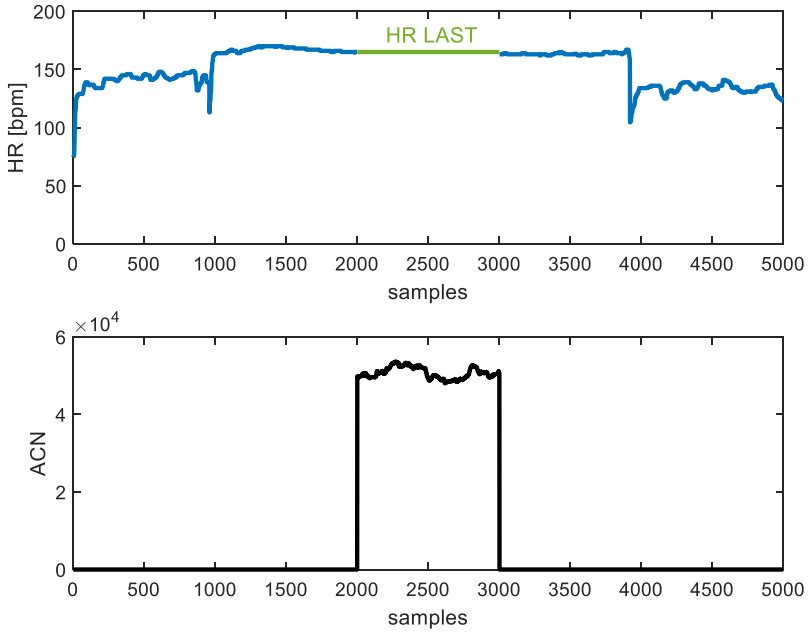


Figure 16: HR LAST method: in the top figure last one value of HR before the corruption starts was substituted over corrupted samples. On the bottom figure, ACN is always the same as was previously explained.

Looking the last one figure, it seems that nothing has changed, HR is close to the HR reference. This happens only because we are considering a long run, where steady state has been reached (both acceleration and heart rate approximately constant) to have a better idea of this method, it's necessary to have a look also in interval trainings.

In *Figure 17*, is showed an interval training. On the top is visible HR in blue line, where the interval for corruption considered starts from 3000 to 5000. In this interval, the substitution of last one value was performed (green line) and is visible how the information about interval

training is lost looking into the dotted blue line. ACN on the bottom figure is the same, zero padding was performed in non-corrupted samples.

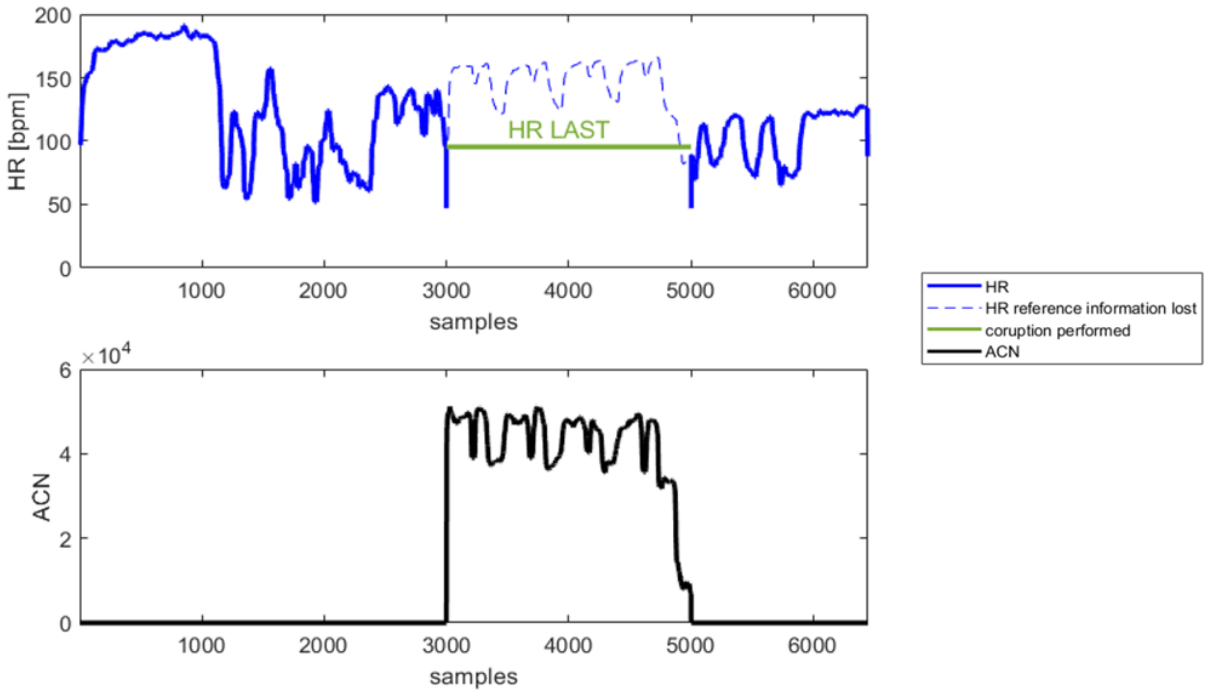


Figure 17: HR LAST method performed on an interval training. On the top is visible HR in blue line, where the interval for corruption considered starts from 3000 to 5000. In this interval, the substitution of last one value was performed (green line) and is visible how the information about interval training is lost looking into the dotted blue line. ACN on the bottom figure is the same, zero padding was performed in non-corrupted samples.

3) HR REST method

Last tested method consists of corrupting the heart rate in the interval selected, with values of HR rest. HR rest is obviously considered for each subject, in this case was considered the value of HR min reached during the night for each subject.

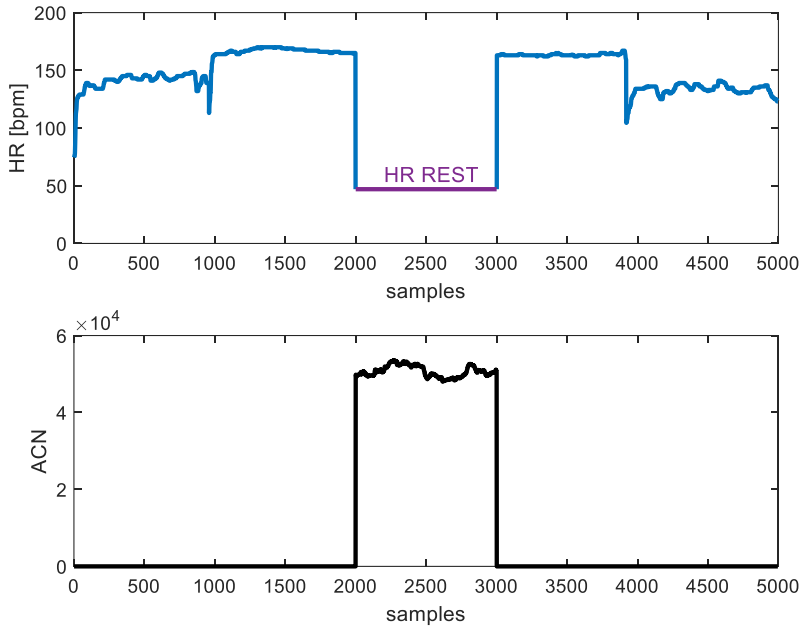


Figure 18: HR REST method: values of HR rest was substituted over corrupted samples, ACN is always the same as was previously explained. HR rest is obviously considered for each subject, in this case was considered the value of HR min reached during the night for each subject.

2.5.3 Test set, training set and cross validation

For training the network, the augmented dataset composed by 176 training sessions, was divided into training set and test set through the leave-one-out method, where one subject was the test set and the other four subjects were the training set. Early stopping was performed considering 30% of training set to be the validation set and the resting 70% was used to train the network.

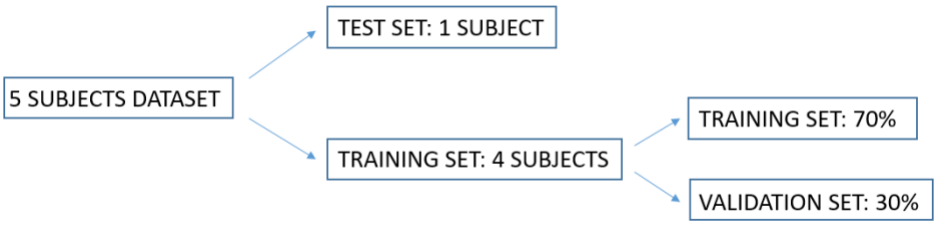


Figure 19: Dataset division into subset

Each set was normalized with the min-max method, obtaining values from 0 to 1, according to:

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad \textit{Equation 18}$$

Where x_n is the normalized value, x is the value to be normalized, x_{min} and x_{max} are the minimum and maximum value of the variable.

Zero padded samples performed during data manipulation, were not included in the normalization.

2.5.4 Architecture and learning parameters

Layer architecture is composed by an input layer with ACN and HR corrupted as input features, an LSTM layer with 32 hidden units, a dropout layer with 45% of probability (in order to avoid overfitting) a fully connected layer and a regression output layer.

Figure 20: Network architecture

Loss function adopted is the half-mean-squared-error loss:

$$loss = \frac{1}{2N}\sum_{i=1}^{N}(t_i - y_i)^2 \qquad \textit{Equation 19}$$

Where *N* is the time series length, t_i is the target output, y_i is the network prediction.

Method used for the stochastic optimization is Adam [50], computationally efficient, requires little memory and works well with long time sequence data.

Training epochs chosen equal to 1000. Learning rate was set constant and equal to 0.01. Validation parameters such us validation frequency equal to 5 epochs and validation patience equal to 15. L2 regularization parameter set to 0.005.

After training the network, the prediction was performed.

2.6 METRICS FOR EVALUATING MODEL PERFORMANCE

The performance of a prediction y_i of a time series, is measured quantifying how well the prediction matches with the test data t_i.

The root mean square error (RMSE) and the mean absolute error (MAE), are commonly used for evaluating the performance of a prediction model.

- The RMSE is calculated as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(t_i - y_i)^2}{N}} \qquad \textit{Equation 20}$$

- The MAE is calculated as:

$$MAE = \frac{\sum_{i=1}^{N}|(t_i - y_i)|}{N} \qquad \textit{Equation 21}$$

Where y_i is the predicted value, t_i is the test value, N is the total length of the signal considered.

These two metrics give a different information in terms of response to outliers. RMSE has the benefit of penalizing large errors more than MAE, but on the other hand RMSE does not describe average error like MAE does. MAE will be always smaller than RMSE.

The metrics above mentioned are the most commonly used.

Another metric useful for comparing performance of predictions methods on time series is the mean percentage error (MPE).

While MAE describes the average magnitude of residuals taking into account only the absolute value, MPE shows how the model predictions are far from the desired output, taking into account both positive and negative errors, allowing to understand if the model is underestimating (negative errors) or overestimating (positive errors).

Is statistic, MPE is the average of percentage errors by a predicted value of a model differs from the real value of the quantity being predicted.

- The MPE is calculated as:

$$MPE = \frac{100\%}{N} \sum_{i=1}^{N} \frac{t_i - y_i}{t_i} \qquad \textit{Equation 22}$$

3 RESULTS AND DISCUSSION

3.1 BETTER METHOD FOR DATA MANIPULATION

Looking into the RMSE, between three methods tested, HR LAST is the one with a smaller RMSE distribution and a lower median.

For this reason, HR LAST was the adopted method for HR prediction through LSTM network.

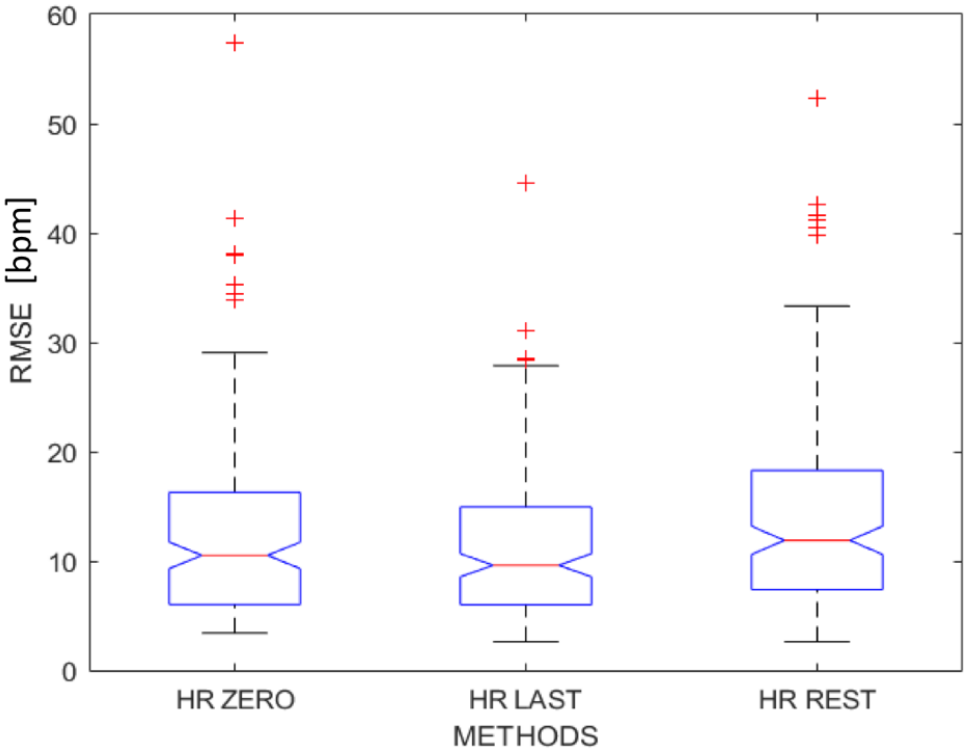


Figure 21: RMSE between three methods tested: HR LAST has smaller RMSE distribution and a lower median.

3.2 PERFORMANCE TRAINING SET AND TEST SET

Evaluating a machine learning model means also to evaluate it through the training set.

As already said, the dataset, without considering cross validation, is split into two subset: a test set and a training set. Performance between them is different. We expect a higher performance for the training set because the net is trained through this, and a larger error for the test set. If the opposite is happening, an overfitting of the network is probably occurring.

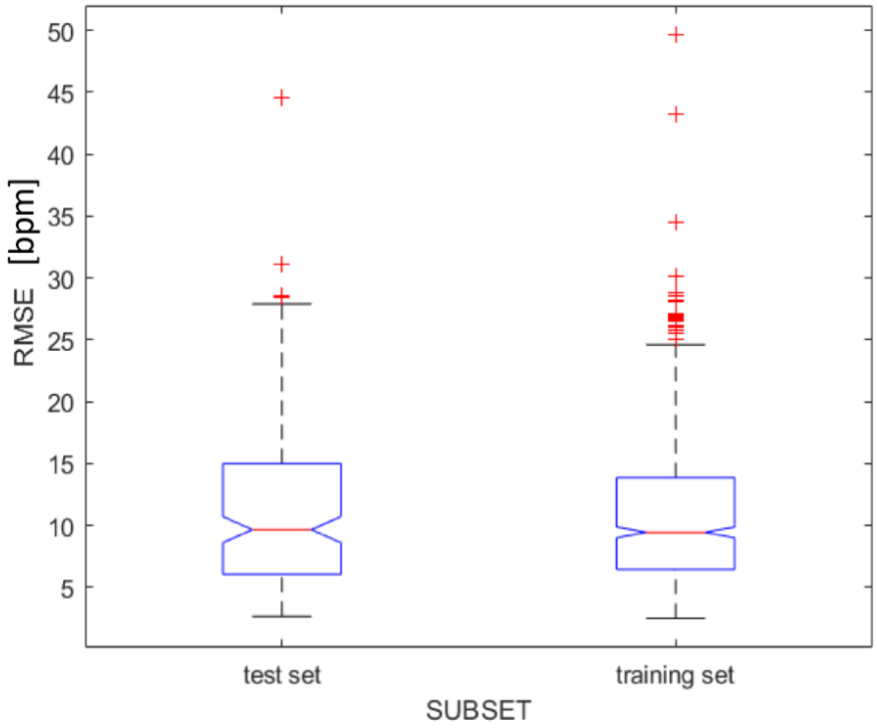


Figure 22: Training set and Test set error distribution to observe network performance and overfitting

An example of prediction between training set and test set is showed in *Figure 23*.

Training set prediction is more accurate compared to that one of test set, the RMSE results to be equal to 6.26 bpm and 8.1 bpm respectively and MAE results to be 4.28 bpm and 5.33 bpm respectively.

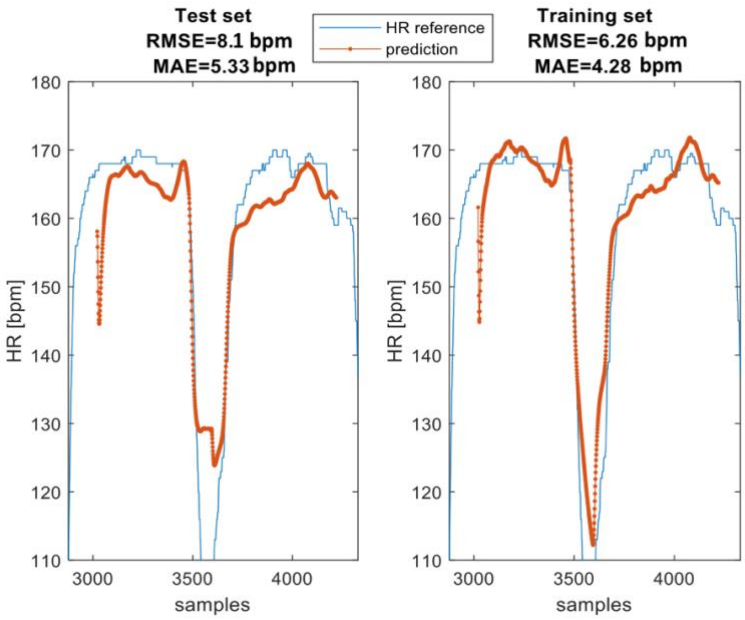


Figure 23: Training set and test set predictions. Training set shows a better prediction with a smaller error because the net has been trained and had learned with this set. If the opposite is happening, an overfitting of the network is probably occurring

Looking into the graph of loss function, no overfitting is occurring. Validation set loss is always lower than training set loss.

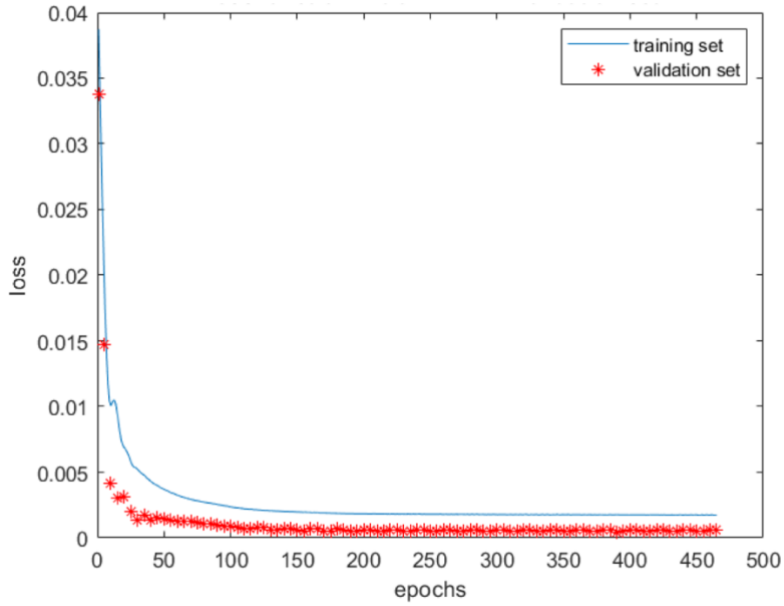


Figure 24: Loss function between validation set and training set to observe that no overfitting is occurring

3.3 LINEAR REGRESSION PREDICTION AND LSTM PREDICTION

With coefficients calculated before by the linear regression, heart rate prediction was made using activity counts.

Performance comparison between predictions made by the LSTM model and by linear regression shows that both MAE and RMSE distribution results with a lower range compared to the linear regression.

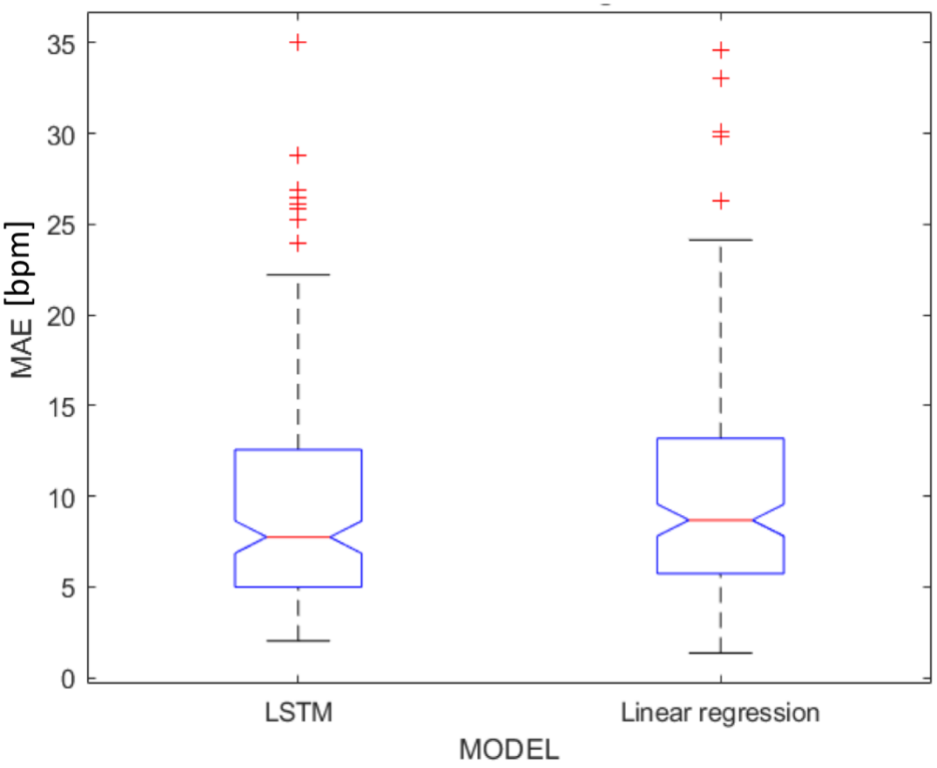


Figure 25: MAE distribution of two methods predictions: the LSTM model and the linear regression model. LSTM error distribution has a lower range of error compared to the linear regression model.

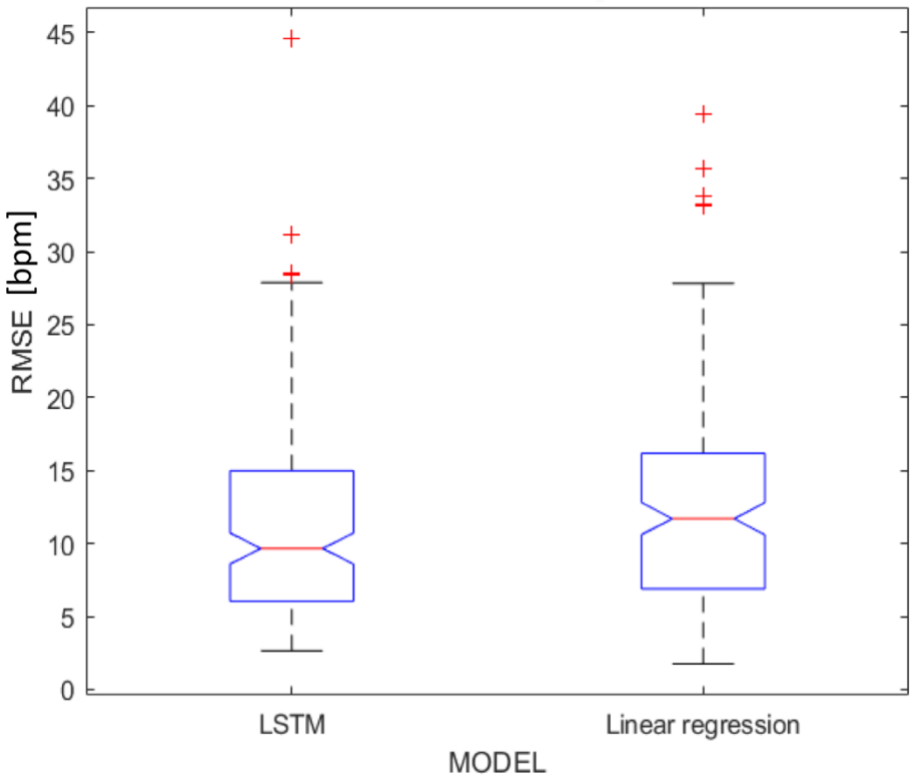


Figure 26: RMSE distribution of two methods predictions: the LSTM model and the linear regression model. LSTM error distribution has a lower range and median value error compared to the linear regression model.

3.3.1 Error in training load intensity zones

As was already mentioned in *section 1* training load intensity zones, calculated according to HR, are five.

For athletes in general, in particular for endurance runners, it is important to have a good HR measurement, in order to estimate training load. For this reason, error is also calculated per training zone. In this case, is important to define an acceptable error for accurate heart rate.

The *American National Standard of Cardiac monitors heart rate meters* defines, for ECG devices monitors, a maximum deviation in amplitude for a time-varying output signal of ±10% from the input [51].

We are not dealing with comparisons between input and output signals from an ECG device, we are considering a prediction on HR signals, but we can consider it as margin of error for this case. Any margin of error has been already defined for this particular case.

Considering heart rate reference for each of five zone, mean percentage error (MPE) has been calculated both for LSTM model and linear regression model.

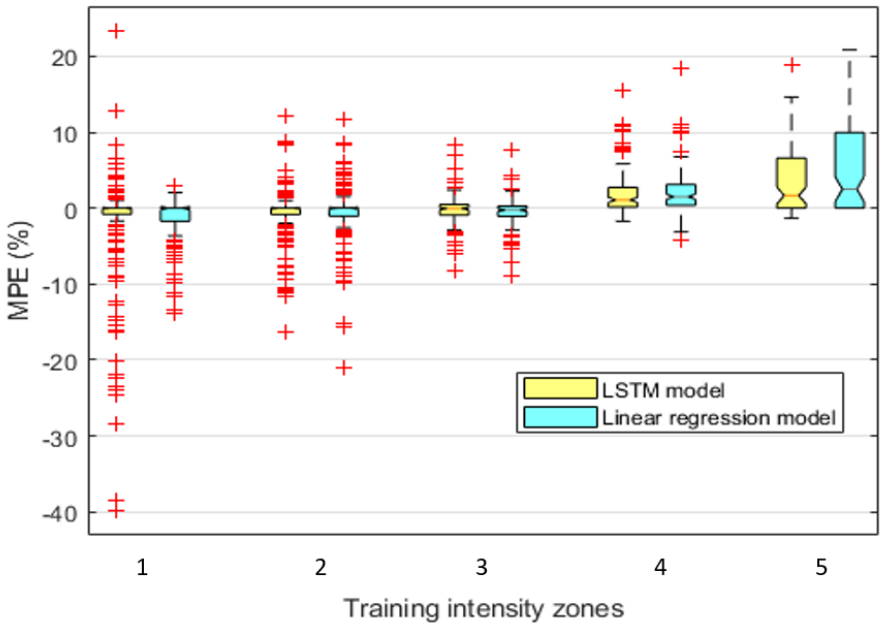


Figure 27: mean percentage error (MPE) calculated in the five HR intensity zones for the LSTM model and linear regression model. MPE in all five zones is inside the range defined (±10%) except zone 5. For the LSTM model until 89^{th} percentile of the error is under the threshold but its maximum value reaches +16.5%, an overestimation from 10% to 16.5% is equal to 22 ± 7 bpm (mean ± standard deviation). For the linear regression model instead, 76^{th} percentile of this distribution is exactly at the limit of the threshold, contrary to LSTM model, reaching its maximum value at 20.85% of MPE equal to 23 ± 16 bpm.

From the distribution of mean percentage error for each zone, MPE in all five zones is inside the range defined (±10%) except zone 5. For the LSTM model until 89^{th} percentile of the error is under the threshold but its maximum value reaches +16.5%, an overestimation from 10% to 16.5% is equal to 22 ± 7 bpm (mean ± standard deviation). For the linear regression model instead, 76^{th} percentile of this distribution is exactly at the limit of the threshold, contrary to LSTM model, reaching its maximum value at 20.85% of MPE equal to 22 ± 17 bpm.

LSTM model was significantly different to the linear regression model: with analysis of variance (ANOVA) performed, $\mathrm{p-value} < 0.05$ considering a significance level of 95% ($\alpha = 0.05$).

Overestimating training intensity zones, in general, means for an athlete to have done more than was actually done in terms of training activity.

In zones below the anaerobic threshold (anaerobic threshold is reached at almost 85% of HRmax [52] [53] , inside zone 4), overestimating heart rate is more critical because the athletes may think to have reached higher intensity. Whereas, this is the result of a not reliable HR. This accuracy may lead to train less or run at sub-optimal speed having lower performance improvements.

Overestimating zone 5, as in this case, could be less of a problem because zone 5 is the maximum intensity zone, which is less frequently used by runners during their trainings.

Zone 5 is a very vigorous intensity training zone, mostly reached during interval trainings and sprint trainings, zone 1 and 2 are the most easy intensity zones, reached mostly at the beginning of every training session or during recovery trainings.

For these reasons, an overestimation or also an underestimation in zone 3 and 4 may be more crucial. This is because are the most frequent zones reached during running.

Figure 28 shows the distribution in training zone duration during running trainings. Blue bars are intensity zone durations according to the HR reference, in orange according to the HR predicted by the LSTM model, in yellow according to the linear regression model.

The duration displayed is cumulative of all the dataset composed by 176 running sessions.

Zone 3 and 4 are the most frequent reached zones during training, overestimating or underestimating HR above the threshold (±10%) in these zones would be more crucial than in zone 1, 2 and 5. Of course, overestimating or underestimating HR, even if HR predicted is inside the threshold range defined (±10%), has repercussions on the duration of each training zone. For example, if HR predicted in zone 1 has been overestimated compared to the HR reference, is probably that these values are now part of zone 2, causing a longer duration of this zone.

Comparing, indeed, HR reference with HR predicted by the LSTM model and by linear regression model, the HR of LSTM model shows quite the same duration as HR reference for zone 1, 2, 4, while HR of linear regression model shows quite the same duration as HR reference only for zone 4. Finally, the two models show almost the same duration for zone 3 and 4.

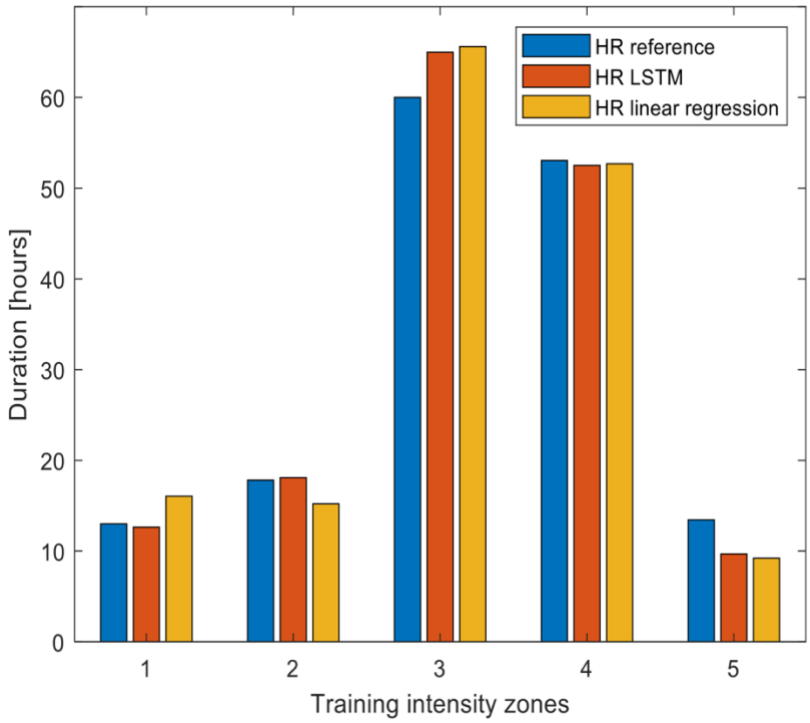


Figure 28: Distribution in training zone duration. In blue bars intensity zone durations according to the HR reference, in orange according to the HR predicted by the LSTM model, in yellow according to the linear regression model. The duration displayed is cumulative of all the dataset composed by 176 running sessions. The two models show almost the same duration for zone 3 and 4.

A variation in training intensity zone duration itself has repercussions on the calculation of training load (see *Equation 1*).

This causes a propagation of errors visible in *Figure 29*, showing a comparison between training loads calculated for each of the 176 running sessions. Training load is displayed in normalized values and has been calculated considering duration in zones of HR reference (green line), duration in zones of HR predicted by the LSTM model (blue star line) and the duration in zones of HR predicted by the linear regression model (red circle line).

The MPE calculated in non-normalized values, between the training load of HR reference and the training load of HR predicted by the two model is equal to 1.12% for the LSTM model and 1.62% for the linear regression model.

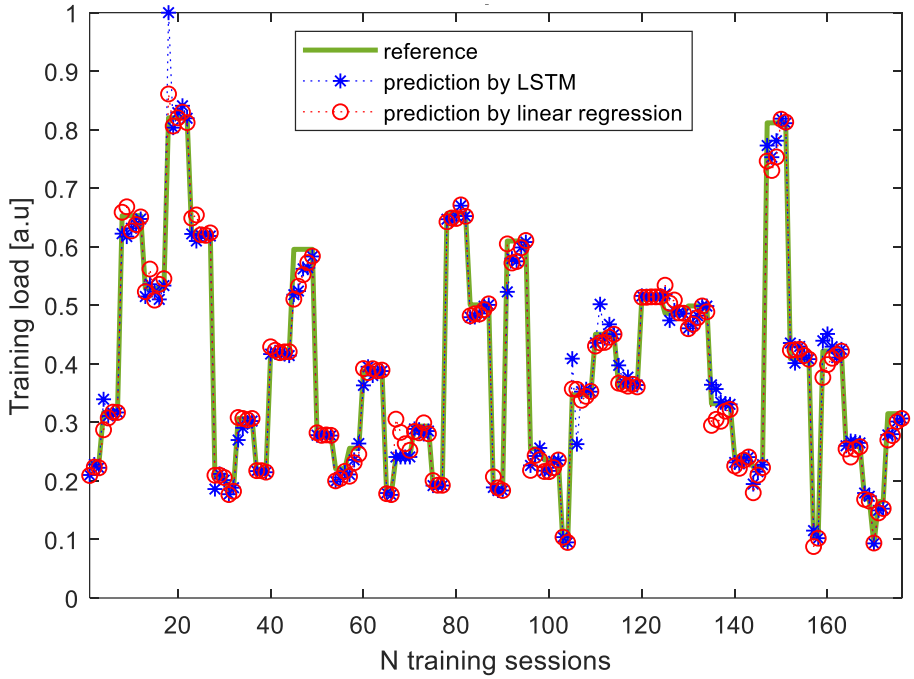


Figure 29: Comparison between training load calculated for each of the 176 running sessions considering duration in zones of HR reference (green line), duration in zones of HR predicted by the LSTM model (blue star line) and the duration in zones of HR predicted by the linear regression model (red circle line). The mean percentage error between the training load of HR reference and the training load of HR predicted by the two model is equal to 1.12% for the LSTM model and 1.62% for the linear regression model.

Some examples of predictions made by LSTM model and linear regression model. The blue line the reference heart rate, in orange line the predicted heart rate by each model with RMSE and MAE above each figure.

In *Figure 30* is clear how in linear regression the different kinetics between ACN and HR were not took into account. At the beginning of the prediction, for the linear regression model, the HR predicted is already at 158 bpm, while for LSTM model is rising from 138 bpm. The linear regression prediction reflects too much the kinetic of ACN (*see Figure 10*). LSTM model, instead, follows more the trend of HR reference. This behaviour is also more clear looking into the periods in which HR decrease and increase again.

Also errors result to be lower for the LSTM model compared to the linear regression. RMSE results to be equal to 5.3 bpm and 8.93 bpm respectively and MAE results to be 4.37 bpm and 7.45 bpm respectively.

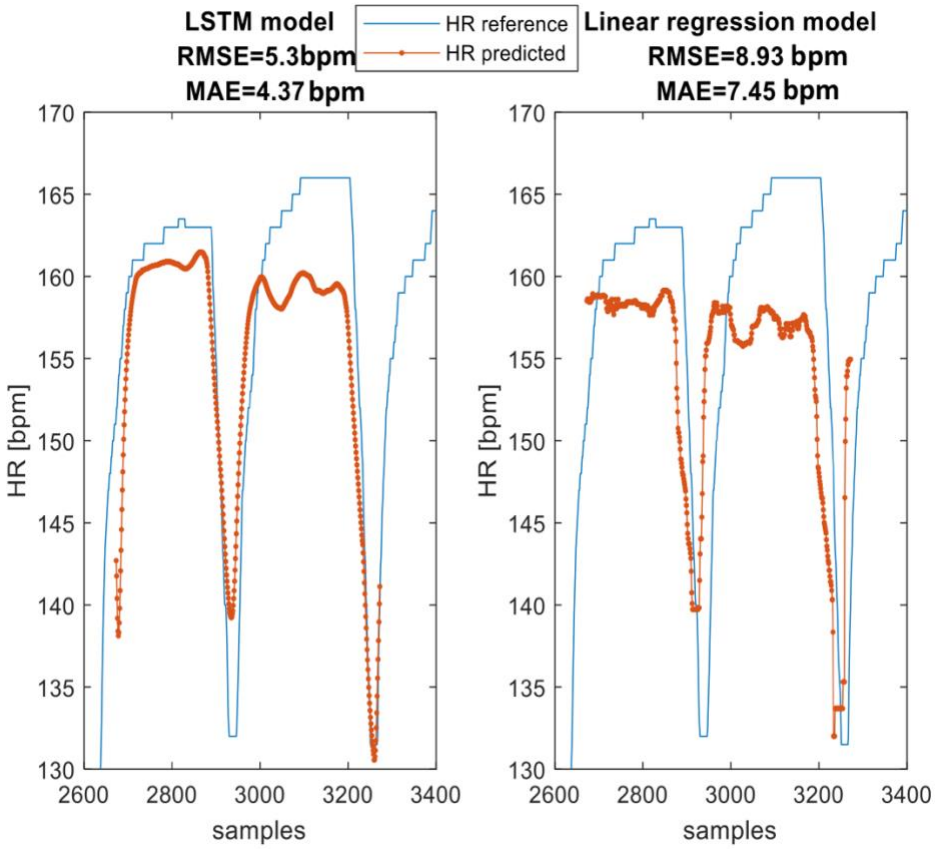


Figure 30: LSTM prediction and linear regression prediction: in linear regression the different kinetics between ACN and HR were not took into account. At the beginning of the prediction, for the linear regression model, the HR predicted is already at 158 bpm, while for LSTM model is rising from 138 bpm. The linear regression prediction reflects too much the kinetic of ACN. LSTM model, instead, follows more the trend of HR reference. This behaviour is also more clear looking into the periods in which HR decrease and increase again.

In ***Figure 31*** is showed a prediction on an interval training. Is visible also in this case, the two different kinetics are not took into account for the linear regression model, while the LSTM prediction follows HR reference timing. Furthermore, at the end of the four intervals, HR decreases in a perfectly linear way for the linear regression model, instead, for the LSTM prediction has an exponential decrease that reflects more the trend of HR reference.

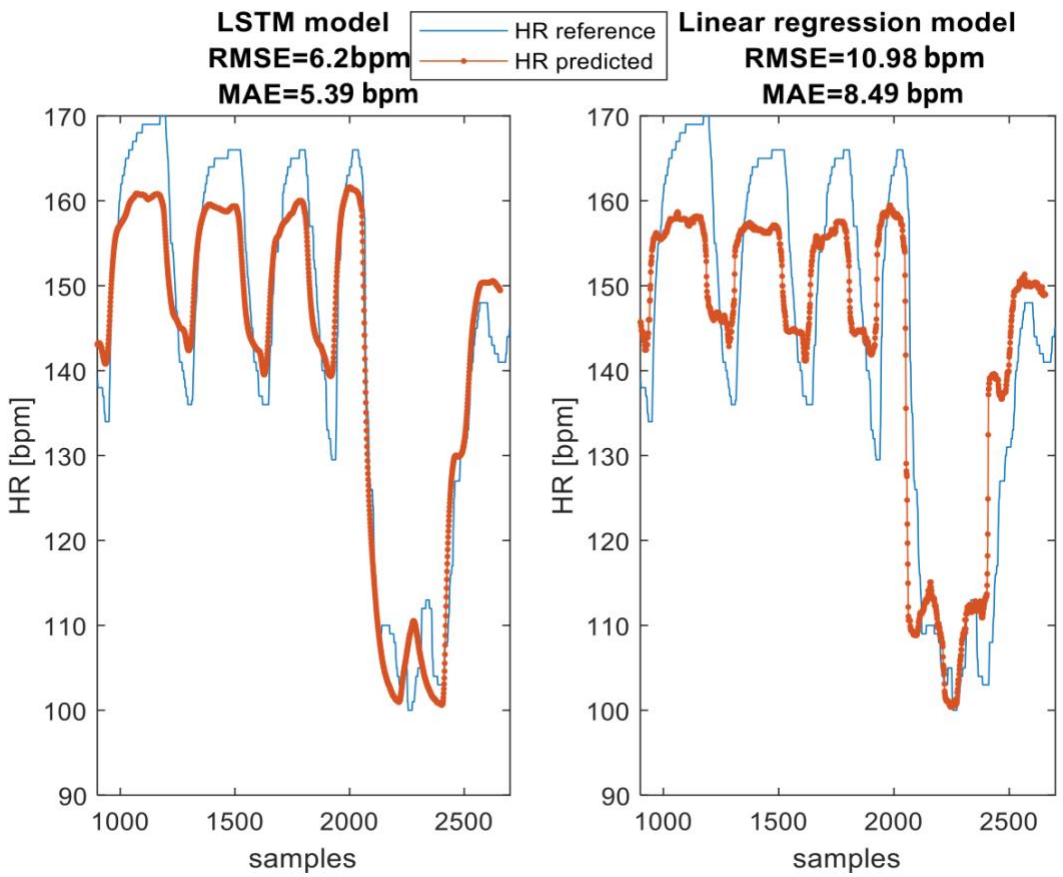


Figure 31: Prediction on an interval training. Is visible also in this case, the two different kinetics are not took into account for the linear regression model, while the LSTM prediction follows HR reference timing. Furthermore, at the end of the four intervals, HR decreases in a perfectly linear way for the linear regression model, instead, for the LSTM prediction has an exponential decrease that reflects more the trend of HR reference.

Figure 32 reflect perfectly the linear relation between ACN and HR considered with the linear regression model against a better timing prediction with LSTM model. While linear regression model has often moments in which the prediction is completely flat due to the subject stop (acceleration constant), the LSTM model follows more heart rate trends that hardly shows a constant behavior. In this case, also half value for both RMSE and MAE of LSTM model is reached compared to the linear regression model.

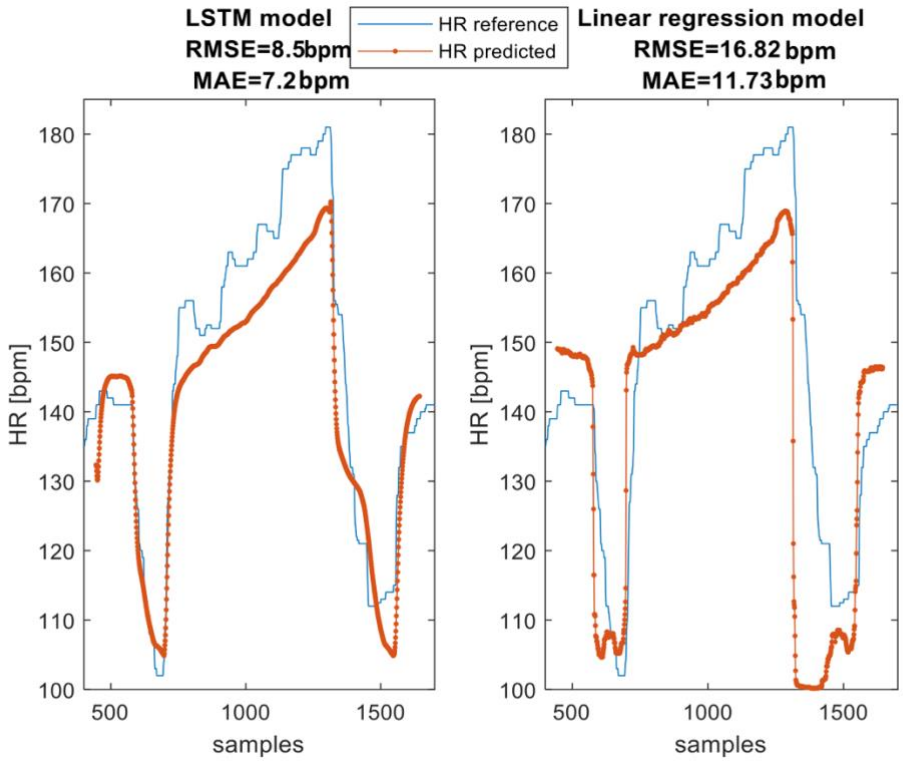


Figure 32: While linear regression model has often moments in which the prediction is completely flat due to the subject stop (acceleration is constant), the LSTM model follows more heart rate trends that hardly have a constant behavior. In this case, RMSE and MAE of the linear regression model is the double compared to the LSTM model.

The choice of an LSTM model for this study was led mainly by these issues reported by linear regression.

Training a network with time sequence data, the network can learn the history information for predicting heart rate during running training sessions, and thanks to the 'long-short-term memory' the information over time is not lost.

3.4 Comparison between HR monitored by OHRM and HR predicted by LSTM

The LSTM network trained with the augmented data was tested on the data recorded by the OHRM with the real corruption. In this way, we can evaluate our solution in a real case scenario.

The starting dataset of 45 running session was composed by 30 HR signals corrupted monitored by the OHRM. The prediction was indeed performed over 30 heart rate signals.

3.4.1 Error in training load intensity zones

In *Figure 33*, to evaluate the overestimation or underestimation in training intensity zones, MPE is considered. Yellow boxplots refer to the predicted HR by the LSTM model, blue boxplots refer to the HR corrupted monitored by OHRM. HR predicted by the LSTM is inside the threshold (±10%) except for zone 2,3,5 whose 98th, 99th, 86th percentiles are inside the threshold but they reach respectively maximum at -11.99%, -10.86%, 11.57%, values very close to the threshold. HR corrupted monitored by the OHRM, instead, is overcoming of a big MPE the threshold in zones 2,3,4,5 respectively of 21.82%, 27.85%, 41.04% 43.79% and how was mentioned in *section 3.3.1*, is more crucial overestimate or underestimate zone 3 and 4 because are the most frequent zones during running. This low-accuracy behaviour of OHRM during increasing workload is already known in literature [18] [3].

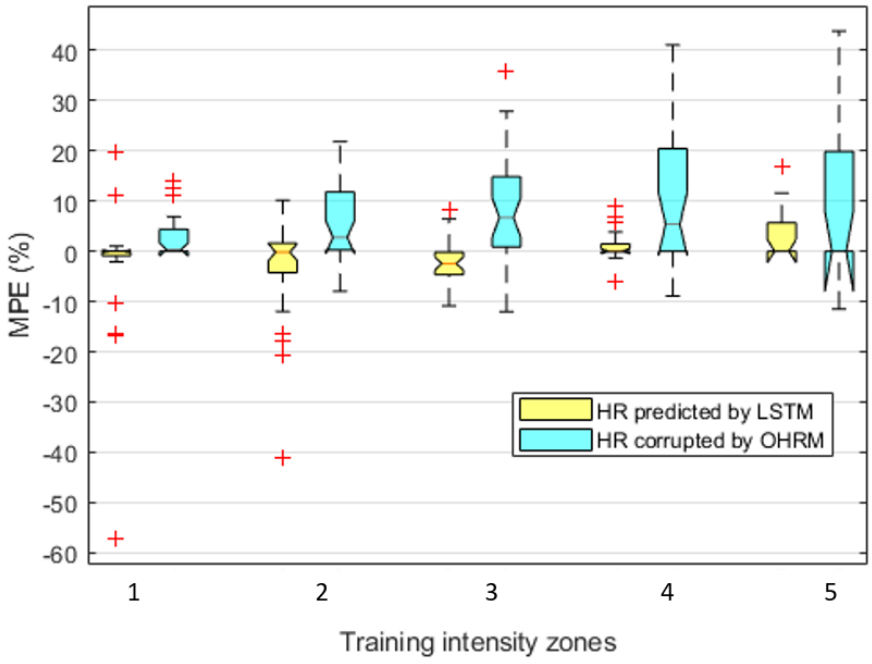


Figure 33: MPE is considered to evaluate the overestimation or underestimation in training intensity zones in particular for HR calculated monitored by the OHRM on PPG signals corrupted. Yellow boxplots refer to the predicted

HR by the LSTM model, blue boxplots refer to the HR corrupted. HR predicted by the LSTM is inside the threshold (±10%) except for zone 2,3,5 whose 98th, 99th, 86th percentiles are inside the threshold but they reach respectively maximum at -11.99%, -10.86%, 11.57%, values very close to the threshold. HR corrupted monitored by the OHRM, instead, is overcoming of a big MPE the threshold in zones 2,3,4,5 respectively of 21.82%, 27.85%, 41.04% 43.79% and how was mentioned in section 3.3.1, is more crucial overestimate or underestimate zone 3 and 4 because are the most frequent zones during running

Also in this case, overestimating or underestimating has repercussions in training intensity zones duration and, as a consequence, also in the calculation on training load.

In *Figure 34* is displayed the duration per intensity zones between HR reference in blue bars, HR predicted by LSTM model in orange bars and in yellow bars the HR corrupted monitored by the OHRM.

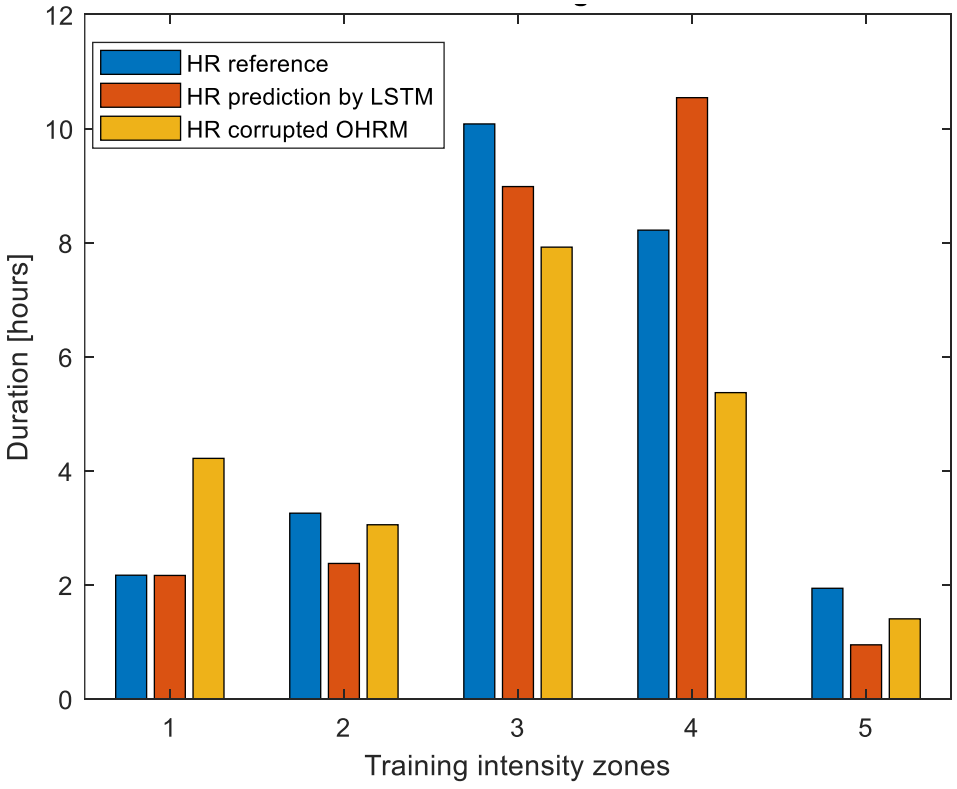


Figure 34: This figure shows the difference in duration per intensity zones. HR reference in blue bars, HR predicted by LSTM model in orange bars and in yellow bars the HR corrupted monitored by the OHRM. Each duration has an impact on the calculation of training load

Each duration has an impact on the calculation of training load (TL) displayed in *Figure 35* . In green line the TL calculated by the HR reference, in blue line TL calculate by the HR predicted by the LSTM model and in red TL calculated by the HR calculated by the OHRM on PPG signals corrupted.

The MPE for training load calculated through the HR predicted by LSTM and the HR corrupted by the OHRM is equal to -0.81% and 19.77% respectively. The difference between these two MPE is big and while for the LSTM model the error is little and may be negligible, the error made by the OHRM is not. A big error in estimating training load for an elite runner could led to risks linked to same athlete's health. Risks can include injuries, overtraining and any performance improvements.

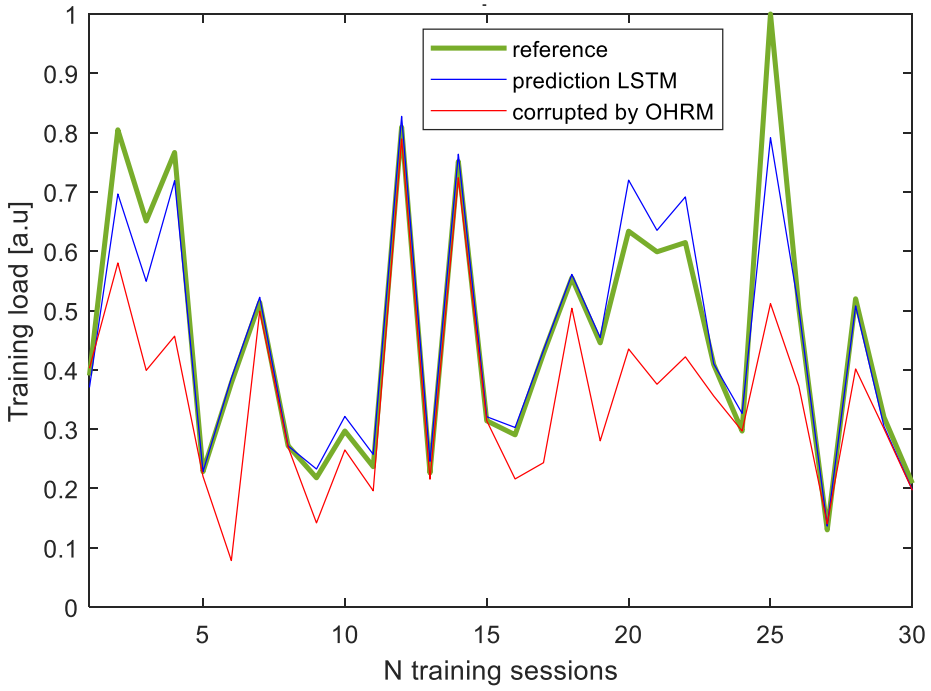


Figure 35: In green line the TL calculated by the HR reference, in blue line TL calculated by the HR predicted by the LSTM model and in red TL calculated by the HR calculated by the OHRM on PPG signals corrupted. The MPE for training load calculated through the HR predicted by LSTM and the HR corrupted by the OHRM is equal to -0.81% and 19.77% respectively. The difference between these two MPE is large and while for the LSTM model the error is little and may be negligible, the error made by the OHRM is big and may be not negligible.

The International Olympic Committee defined that there is a scientific evidence in the relationship between training load and health [54]. Training with a high intensity maximise the performance but if is repeated for a long time it can be dangerous for the athlete's health. The same is in case of poor load that, for elite athletes, may increase factors for injuries [54] [55].

For this reason, the estimation of the training load is very important and making a big error could be dangerous. If TL is overestimated, the athlete may train less in order to recover, leading the runner to have any improvements in terms of physical performance as well as being injured. If, instead, the training load is underestimated, the athlete may led to train with a higher load with injury and overtraining consequences.

Figure 36, Figure 37, Figure 38 show predictions made by LSTM over HR corrupted by the OHRM. In blue line the reference HR, in orange the predicted HR by the LSTM and in black line the corrupted HR.

Above each figure, the RMSE evaluated both for HR predicted and HR corrupted.

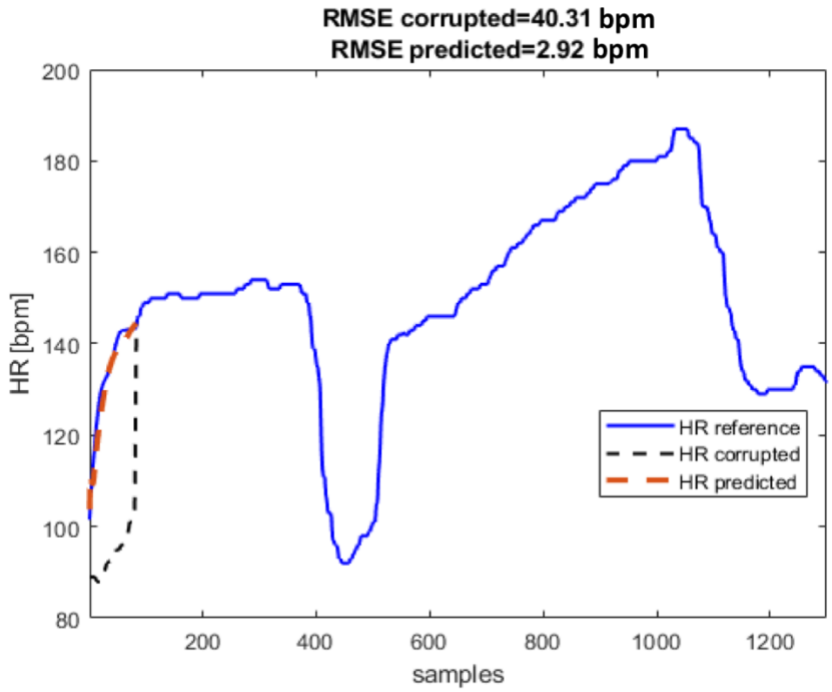


Figure 36: HR during VO2max test. In blue line the reference HR, in orange the predicted HR by the LSTM and in black line the corrupted HR. RMSE is improved from 40.31 bpm to 2.92 bpm

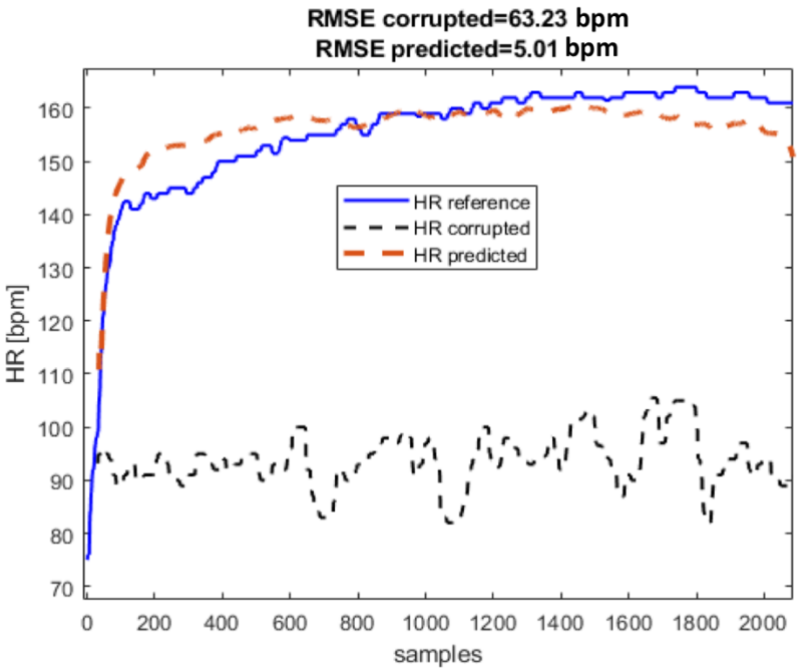


Figure 37. HR during a long distance run: in blue line the reference HR, in orange the predicted HR by the LSTM and in black line the corrupted HR. RMSE is improved from 63.23 bpm to 5.01 bpm

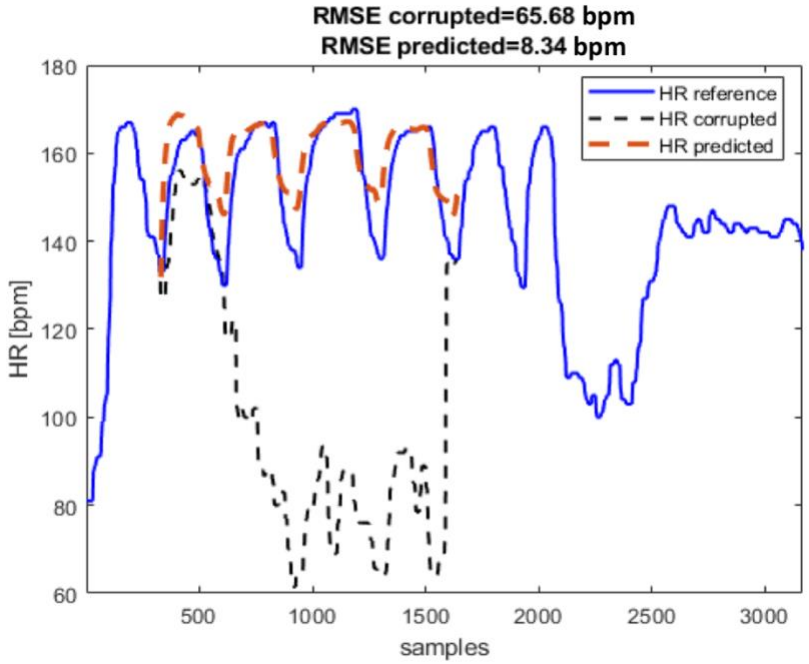


Figure 38: HR during am interval training. In blue line the reference HR, in orange the predicted HR by the LSTM and in black line the corrupted HR. RMSE is improved from 65.68 bpm to 8.34 bpm

3.5 LITERATURE COMPARISON

Jindal et al [20], using a wrist-worn watch with a PPG and a tri-axial accelerometer, built a deep learning classification model for determining heart rate during low and high intensity exercise in 15 males, using heart rate features, where motion data were used only to filter the signal before input the heart rate features in the model. We cannot compare our LSTM regression model with this classification model in term of accuracy, the only error reported by Jindal et al. [20] is an average error in 140s window length of prediction equal to 5.72% calculated during a high intensity exercise.

Considering a window length of 140s in random samples for each prediction made by this study with LSTM model, the average value of mean percentage error is 0.95%.

	MPE study [20]	MPE this study
140s HR prediction	5.72%	0.95%

Table 3: Comparison between this study and study [20] *using a deep learning model for HR prediction using HR features*

Ming and Jun [22] built a classification feedforward neural network using acceleration data and heart rate data for heart rate prediction in 90 minutes signal from a healthy male. Heart rate was recorded during daily life through a portable HR monitor with electrodes and accelerometer data were recorded through a tri-axial accelerometer. The author reported a value of mean absolute error for 30s prediction on training set and on test set respectively equal to 3.12 bpm and 3.31 bpm.

As was done for Jindal et al. study [20], considering a window length of 30s in random samples for each prediction made by training set and test set of this study with LSTM model, median MAE results to be equal to 6.81 bpm and 7.56 bpm, minimum MAE instead results to be 0.21 bpm and 0.43 bpm respectively for training set and test set. The Median MAE was higher compared to the

study [22] but is not clear if the value reported by the author of [22] was a minimum, an average, a median or value of MAE.

30s HR prediction	MAE study [22]	Median MAE this study	Minimum MAE this study
Training set	3.12 bpm	6.81 bpm	0.21 bpm
Test set	3.31 bpm	7.56 bpm	0.43 bpm

Table 4: Comparison between this study and study [22] *using a feedforward NN.*

Is not exactly clear what MAE the author reported, median MAE in our study result to be higher and minimum MAE results to be lower.

3.6 LIMITATIONS

1) Cardiovascular Drift

In *section 2.4 Figure 10,* cardiovascular drift has been mentioned.

In *Figure 39* our LSTM model does not consider a cardiovascular drift lasting for 16 minutes, predicting a flat line. A reason because the algorithm is not able to correctly predict the HR, is because is the only training session with a long cardiovascular drift duration in the dataset. Even with the data augmentation, the availability of training of this type is not enough to predict it correctly. More trainings with a big cardiovascular drift in the dataset can allow to the network to learn the trend predicting it in a better way.

Although the prediction of HR is mostly flat, the error (RMSE=9 bpm, MAE=7.08 bpm) it is not extremely high, and especially the MPE (MPE=1.17%) is inside the range previously defined ($\pm 10\%$).

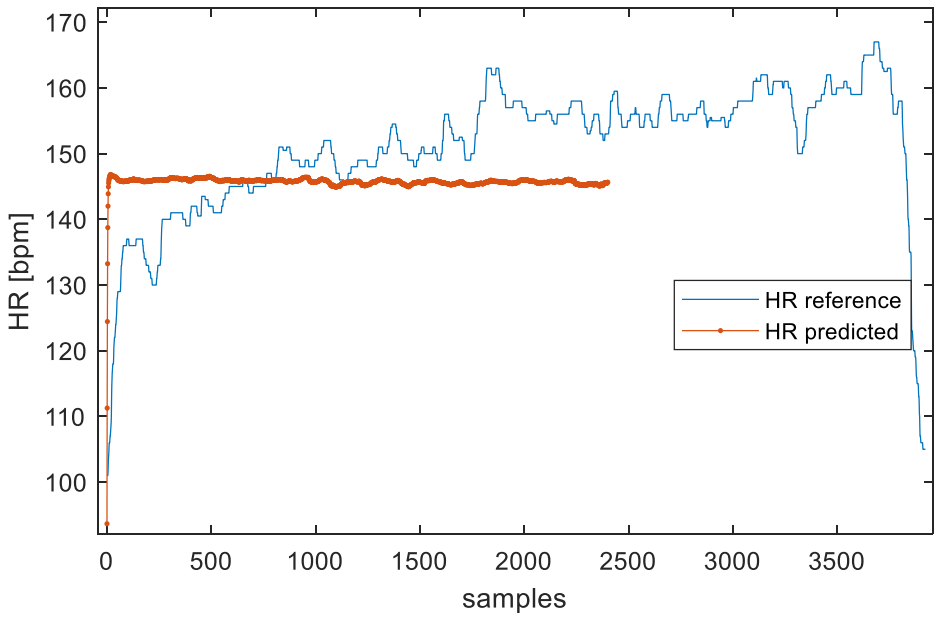


Figure 39: the LSTM model does not consider a cardiovascular drift lasting for 16 minutes, predicting a flat line. A reason because the algorithm is not able to correctly predict the HR, is because is the only training session with a long cardiovascular drift duration in the dataset. Even with the data augmentation, the availability of training of this type is not enough to predict it correctly. More trainings with a big cardiovascular drift in the dataset can allow to the network to learn the trend predicting it in a better way. Although the prediction of HR is mostly flat, the error (RMSE=9 bpm, MAE=7.08 bpm) it is not extremely high, and especially the MPE (MPE=1.17%) is inside the range previously defined ($\pm 10\%$).

2) PARTS in which the subject is not running.

It is possible that HR is also corrupted in periods in which the subject is not running but is still in movement. These parts involved are usually two, between warming up and training and also between training and cooling down. During these periods, LSTM model could predict heart rate in a wrongly way.

For example, considering a running session in *Figure 40*, where the blue line is the reference HR, the orange line is the predicted HR, is possible to distinguish four parts, a warming up (1), a window between the warming up and the training (2), an interval training (3) and a cooling down (4) and the prediction is made in the second part.

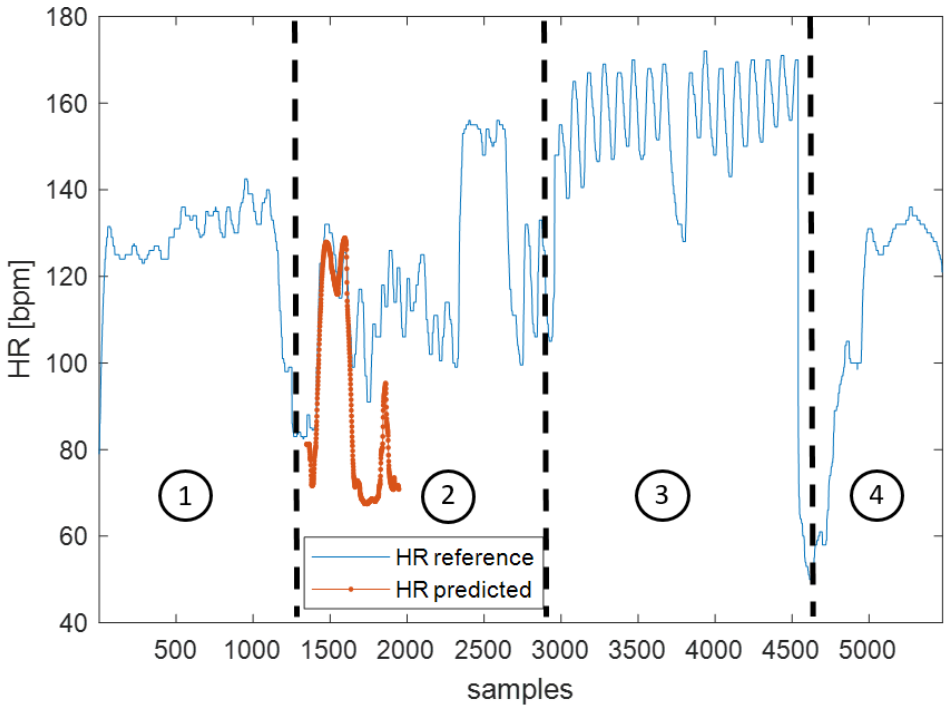


Figure 40: the blue line is the reference HR, the orange line is the predicted HR, is possible to distinguish four parts, a warming up (1), a window between the warming up and the training (2), an interval training (3) and a cooling down (4) and the prediction is made in the second part.

In *Figure 41* prediction is showed closer. From the beginning until sample 1650 the prediction is really close to the reference heart rate, but after this period, the prediction is wrong. During

periods between warming up and training and also between training and cooling down, runners usually do some stretching exercise, exercises with arms or legs, and so on.

Over all, the prediction is 50% good and 50% wrong, this allow to have a big error in term of RMSE, MAE and MPE (RMSE=27.52 bpm, MAE=21.91 bpm, MPE=19%). The reason is because the subject was running during the correct prediction (we can see that from 80 bpm HR rises reaching 130 bpm) and was not running during the wrong prediction (beats are still quite high because the subject just stopped running but they tend to decrease).

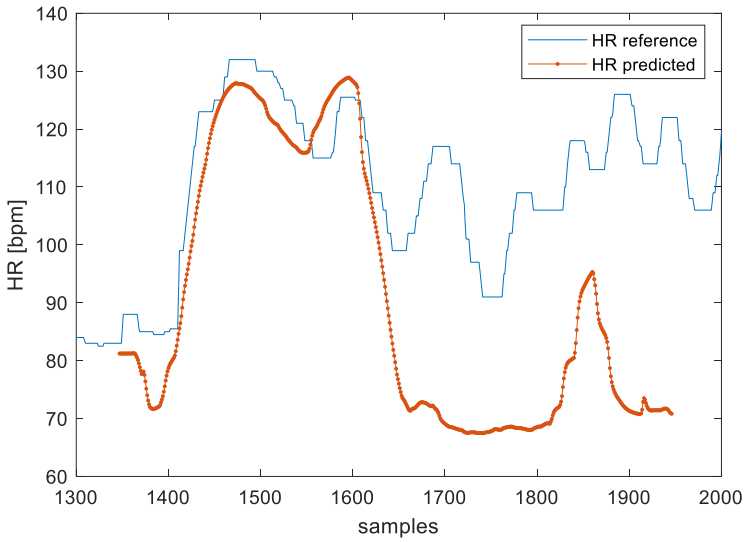


Figure 41: From the beginning until sample 1650 the prediction is really close to the reference heart rate, but after this period, the prediction is weak. During periods between warming up and training and also between training and cooling down, runners usually do some stretching exercise, exercises with arms or legs, etc… Over all, the prediction is 50% good but 50% wrong, this allow to have a big error in term of RMSE, MAE and MPE. The most probable reason is because the subject was running during the correct prediction and was not running during the wrong prediction.

In *Figure 42* HR reference in blue, HR predicted in orange and ACN in green line. It's clearly visible that the prediction is following ACN trends. Calculating correlation coefficients between HR reference and ACN in the two prediction parts we can say if the subject was running or not because Pearson correlation coefficient is high during running periods due to the high correlation between ACN and HR. The two parts prediction have been considered separately by the red point in the figure.

Pearson correlation coefficient for the first part of the prediction and the second part results to be equal to 0.78 and 0.43 respectively. From these coefficients, it's highly probable that the subject was running in the first part where ACN and HR are strongly related (r=0.78) and in the second part the probability that the subject is running is less due to the low moderate correlation coefficient (0.43).

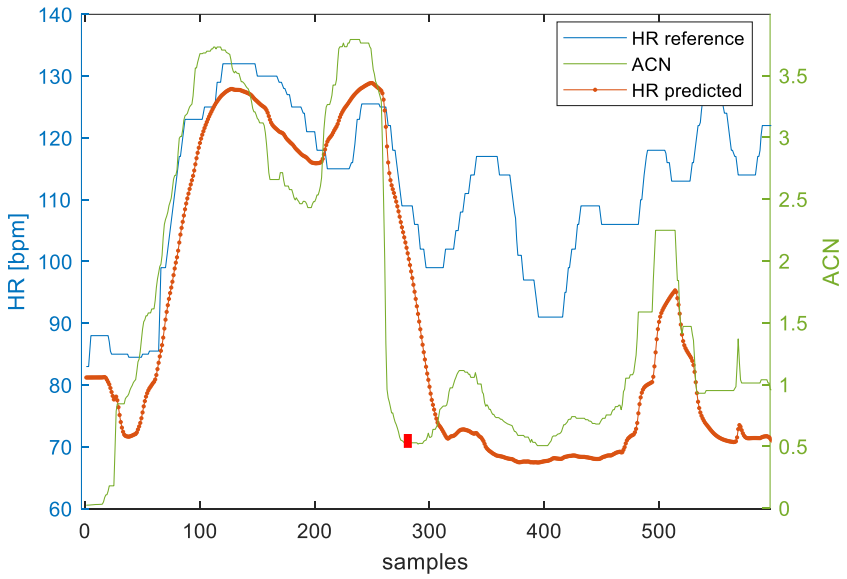


Figure 42: It's clearly visible that the prediction is following ACN trends. Pearson correlation coefficient have been calculated for the two prediction parts considered separately by the red point. It's highly probable that the subject was running in the first part where ACN and HR are strongly related (r=0.78) and in the second part the probability that the subject is running is less due to the low moderate correlation coefficient (0.43).

Not running has a strong impact in the LSTM model, because this algorithm focus on running periods. An activity classifier (based for instance on accelerometer data [56]) can help to separate the type of activity and future research could work on have different LSTM for different activity.

4 CONCLUSIONS

This study presents a novel approach using deep learning technique such us LSTM to predict heart rate during running trainings, using accelerometer features, where heart rate was affected by a low SNR due to the low accuracy of an OHRM.

In general, metrics error of this study model such us MAE, RMSE and MPE were lower than the linear regression model. Both MAE and RMSE have a shorter range of error compared to linear regression prediction.

From the distribution of mean percentage error for each training intensity zone, MPE in all five zones is inside the range defined (±10%) except zone 5 for both models. For the LSTM model until 89^{th} percentile of the error is under the threshold but its maximum value reaches +16.5% (equal to 22 ± 7 bpm). For the linear regression model instead, 76^{th} percentile of this distribution is exactly at the limit of the threshold, contrary to LSTM model, reaching its maximum value at 20.85% (equal to 22 ± 17 bpm).

Overestimating zone 5, is less of a problem because zone 5 is the maximum intensity zone, which is less frequently used by runners during their trainings. Zone 5 is a very vigorous intensity training zone, mostly reached during interval trainings and sprint trainings, zone 1 and 2 are the most easy intensity zones, reached mostly at the beginning of every training session or during recovery trainings. For these reasons, an overestimation or also an underestimation in zone 3 and 4 may be more crucial. This is because are the most frequent zones reached during running.

This little overestimation and underestimation led to a propagation of error for the Edward's training load calculation, where it results to be estimated with a MPE of 1.12% for this study model against 1.62% of the linear regression. In literature, no threshold of error over training load has been defined, but considering an error of 1.12% over this estimation, training load calculated through the HR predicted may be acceptable.

Furthermore, comparing the predicted HR and the HR monitored by the OHRM, the overestimation or underestimation in training intensity zones for the LSTM model results to be inside the threshold (±10%) except for zone 2,3,5 whose 98^{th}, 99^{th}, 86^{th} percentiles are inside the

threshold reaching respectively maximum value at -11.99%, -10.86%, 11.57%, values very close to the threshold. HR corrupted by the OHRM, instead, is overcoming of a big MPE the threshold in zones 2,3,4,5 respectively of 21.82%, 27.85%, 41.04% 43.79% and is more crucial involving zone 3 and 4 because are the most frequent reached zones during running.

Comparing training load estimation calculated through the HR predicted by LSTM and the HR monitored by the OHRM the MPE is equal to -0.81% and 19.77% respectively. The difference between these two MPE is large and while for the LSTM model the error is little and may be negligible, the error made by the OHRM is not. A big error in estimating training load for an elite runner could led to risks linked to same athlete's health. Risks can include injuries, overtraining and any performance improvements.

Future work for improving model performance can be involved. Adding new features like subject-dependent performance features (such us amount of oxygen uptake during each training session, fit index, etc..) may be helpful, because the model of this study only uses motion data (activity counts) and corrupted heart rate as a features. Also a bigger dataset could be really useful for such study comprising more subjects and more running trainings per subject with also an internal variability of the different type of trainings (i.e interval trainings, long distance run, etc..) allowing the network to learn better trends and to predict in a better way cardiovascular drift, overcoming a limitation of this study. Another aspect that can be improved is to understand better the relationship between activity counts and heart rate to overcome a bad prediction in non-running periods also through the help of an activity classifier.

5 APPENDIX

Figure 43: Morning diary, for having info about daily life's athlete (filled by subjects every morning)

MORNING DIARY											
	0	1	2	3	4	5	6	7	8	9	10
Recovery* status right now on a scale from 0 to 10 (10 being fully recovered):											
Physical Condition right now on a scale from 0 to 10 (10 being the best condition)											
Muscle soreness right now on a scale from 0 to 10 (10 being maximal muscle soreness)											
Mental energy right now on a scale from 0 to 10 (10 being maximal energy)											
Physical stamina on a scale from 0 to 10 (10 being maximal stamina)											
Stress level right now on a scale from 0 to 10 (10 being maximal stress)											
	YES						NO				
Are you sick (ill/unwell) right now? yes/no											
Did you travel yesterday? yes/no											
Did you trained the day before? yes/no											
Did you consume alcohol yesterday? yes/no											
Did you watch TV or looked** at other light emitting devices (e.g. smart phone, tablet) shortly (less than an hour) before going to bed, last night? yes/no											
Did you fall asleep watching TV or other light emitting devices, last night? yes/no											
Did you read a book just before falling asleep, last night? yes/no											
What time were you in bed with the intention to sleep (no media, no books, no other…)?											

Figure 44: Training diary, to get info about training activity (filled by subjects every training session)

TRAINING DIARY	
TYPE OF ACTIVITY (e.g. easy run; 5-2-1 interval training)	
DISTANCE (km)	
DURATION (min)	
AVERAGE PACE (bpm)	
Rating of perceived exertion of the training session (see Borg scale below)	
Training motivation when you started the training (in a scale from 0 to 10 where 10 is highest motivation)	
Mental energy level when you started the training (in a scale from 0 to 10 where 10 is the highest energy value)	

Figure 45: Recovery questionnaire, to get info about their recovery status (filled at the end of each week)

RECOVERY QUESTIONNAIRE							
	0	1	2	3	4	5	6
How much effort was required to complete your workouts last week? (in a scale from 0 to 6: excessive effort 0 – hardly any effort 6)							
How recovered did you feel prior to the workouts last week? (in a scale from 0 to 6: still not recovered 0 – feel energized and recharged 6)							
How successful were you at rest and recovery activities last week? (in a scale from 0 to 6: not successful 0 – successful 6)							
How well did you recover physically last week? (in a scale from 0 to 6:never 0 – always 6)							
How satisfied and relaxed were you as you fell asleep in the last week? (in a scale from 0 to 6: never 0 – always 6)							
How much fun did you have last week? (in a scale from 0 to 6: never 0 – always 6)							
How convinced were you that you could achieve your goals during performance last week? (in a scale from 0 to 6: never 0 – always 6)							

6 BIBLIOGRAPHY

[1] F. Sartor, J. Gelissen, R. van Dinther, D. Roovers, G. B. Papini, and G. Coppola, "Wrist-worn optical and chest strap heart rate comparison in a heterogeneous sample of healthy individuals and in coronary artery disease patients," *BMC Sports Sci. Med. Rehabil.*, vol. 10, no. 1, 2018.

[2] J. Achten and A. E. Jeukendrup, "Maximal Fat Oxidation during Exercise in Trained Men," *Int. J. Sports Med.*, vol. 24, no. 8, pp. 603–608, 2003.

[3] T. Schack, C. Sledz, M. Muma, and A. M. Zoubir, "A new method for heart rate monitoring during physical exercise using photoplethysmographic signals," *2015 23rd Eur. Signal Process. Conf. EUSIPCO 2015*, no. December 2016, pp. 2666–2670, 2015.

[4] M. Haddad, G. Stylianides, L. Djaoui, A. Dellal, and K. Chamari, "Session-RPE Method for Training Load Monitoring: Validity, Ecological Usefulness, and Influencing Factors.," *Front. Neurosci.*, vol. 11, no. November, p. 612, 2017.

[5] P. C. Bourdon *et al.*, "Monitoring Athlete Training Loads: Consensus Statement.," *Int. J. Sports Physiol. Perform.*, vol. 12, no. Suppl 2, pp. S2161–S2170, 2017.

[6] T. Stoggl and T. Wunsch, "Marathon running: Physiology, psychology, nutrition and training aspects," *Marathon Run. Physiol. Psychol. Nutr. Train. Asp.*, pp. 1–171, 2016.

[7] B. Scott, R. Lockie, and T. Knight, "A Comparison of Methods to Quantify the In-Season Training Load of Professional Soccer Players.," *… J. Sport. …*, no. March, pp. 195–202, 2012.

[8] A. A. T. S. Canlan, N. E. A. L. W. En, and P. A. S. T. Ucker, "T r b i e t l m d b t," vol. 28, no. 9, pp. 2397–2405, 2014.

[9] T. O. Bompa, *Periodization: theory and methodology of training.* 2009.

[10] G. Valenti and K. R. Westerterp, "Optical heart rate monitoring module validation study," *Dig. Tech. Pap. - IEEE Int. Conf. Consum. Electron.*, no. November, pp. 195–196, 2013.

[11] J. Allen, "Photoplethysmography and its application in clinical physiological measurement.," *Physiol. Meas.*, vol. 28, no. 3, pp. R1-39, 2007.

[12] F. Sartor, G. Papini, L. G. Elisabeth Cox, and J. Cleland, "Methodological shortcomings of wrist-worn heart rate monitors validations," *J. Med. Internet Res.*, vol. 20, no. 7, pp. 1–6, 2018.

[13] H. Han and J. Kim, "Artifacts in wearable photoplethysmographs during daily life motions and their reduction with least mean square based active noise cancellation method," *Comput. Biol. Med.*, vol. 42, no. 4, pp. 387–393, 2012.

[14] B. Sañudo, M. De Hoyo, A. Muñoz-López, J. Perry, and G. Abt, "Pilot Study Assessing the Influence of Skin Type on the Heart Rate Measurements Obtained by Photoplethysmography with the Apple Watch," *J. Med. Syst.*, vol. 43, no. 7, p. 195, 2019.

[15] Wikipedia," Long/ short-term/ memory" , https://en.wikipedia.org/wiki/Long_short-term_memory, June 07, 2019

[16] R. Yousefi, M. Nourani, S. Ostadabbas, and I. Panahi, "A motion-tolerant adaptive algorithm for wearable photoplethysmographic biosensors," *IEEE J. Biomed. Heal. Informatics*, vol. 18, no. 2,

pp. 670–681, 2014.

[17] S. S. Chowdhury, R. Hyder, M. S. Bin Hafiz, and M. A. Haque, “Real-Time Robust Heart Rate Estimation from Wrist-Type PPG Signals Using Multiple Reference Adaptive Noise Cancellation,” *IEEE J. Biomed. Heal. Informatics*, vol. 22, no. 2, pp. 450–459, 2018.

[18] M. Boloursaz Mashhadi, E. Asadi, M. Eskandari, S. Kiani, and F. Marvasti, “Heart Rate Tracking using Wrist-Type Photoplethysmographic (PPG) Signals during Physical Exercise with Simultaneous Accelerometry,” *IEEE Signal Process. Lett.*, vol. 23, no. 2, pp. 227–231, 2015.

[19] Y. Ye, W. He, Y. Cheng, W. Huang, and Z. Zhang, “A robust random forest-based approach for heart rate monitoring using photoplethysmography signal contaminated by intense motion artifacts,” *Sensors (Switzerland)*, vol. 17, no. 2, 2017.

[20] V. Jindal, “MobileSOFT: U: A Deep Learning Framework to Monitor Heart Rate During Intensive Physical Exercise,” *Unknown?*, 2016.

[21] R. McConville *et al.*, “Online Heart Rate Prediction using Acceleration from a Wrist Worn Wearable,” 2018.

[22] Y. Ming and J. Jun, “Heart rate prediction based on physical activity using feedforward neural network,” *Proc. - 2008 Int. Conf. Converg. Hybrid Inf. Technol. ICHIT 2008*, pp. 344–350, 2008.

[23] A. Intelligence and P. Mccorduck, “History of artificial intelligence - Wikipedia.” pp. 1–61.

[24] F. Rosenblatt, “Arlington Hall Station,” *Zhurnal Prikl. Mekhaniki i Tec.*, 1962.

[25] D. Hsu, “Multi-period Time Series Modeling with Sparsity via Bayesian Variational Inference,” 2017.

[26] S. T. For, A. Mathematics, and F. Engineering, “Predicting periodic and chaotic signals using Wavenets.”

[27] Wikipedia, “ Perceptron”, https://nl.wikipedia.org/wiki/Perceptron, February 27, 2018

[28] S. Russell and P. Norvig, *Artificial Intelligence A Modern Approach Third Edition*. 2010.

[29] D. Wang and G. Hinton, “Unsupervised learning: Foundations of neural computation,” *Comput. Math. with Appl.*, vol. 38, no. 5–6, p. 256, 2003.

[30] G. Lewicki and G. Marino, “Approximation of functions of finite variation by superpositions of a sigmoidal function,” *Appl. Math. Lett.*, vol. 17, no. 10, pp. 1147–1152, 2004.

[31] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning Internal Representations Error Propagation - ICS Report 8506,” *Cogn. Sci.*, no. V, 1986.

[32] G. Balestra, “Classificazione e Interpretazione Dati Biomedici Classification.”

[33] Y. Yao, L. Rosasco, and A. Caponnetto, “Dropout: A simple way to prevent neural networks from overfitting” *Journal of Machine Learning Research* vol. 15, pp. 1929–1958, 2014.

[34] N. Srivastava, G. Hinton, “Lepton spectra as a measure of b quark polarization at LEP,” *Phys. Lett. B*, vol. 299, no. 3–4, pp. 345–350, 1993.

[35] J. Schmidhuber and S. Hochreiter, “LSTM can solve hard long time lag problems,” *Adv. Neural Inf.*

Process. Syst. 9, vol. 9, p. 473, 1997.

[36] T. U. M. Sepp Hochreiter and I. Jurgen Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[37] H. Tanaka, K. D. Monahan, and D. R. Seals, "Age-predicted maximal heart rate revisited," Elsevier Masson SAS, 2001.

[38] F. Sartor *et al.*, "Estimation of maximal oxygen uptake via submaximal exercise testing in sports, clinical, and home settings," *Sport. Med.*, vol. 43, no. 9, pp. 865–873, 2013.

[39] A. G. Bonomi, *Physical activity recognition using a wearable accelerometer New perspectives for energy expenditure assessment and health promotion*, vol. 41, no. 9. 2009.

[40] Wikipedia, " Linear regression", https://en.wikipedia.org/wiki/Linear_regression, June 15, 2019

[41] Wikipedia, " Pearson correlation coefficient", https://en.wikipedia.org/wiki/Pearson_correlation_coefficient, June 19, 2019

[42] A. V. Kharshikar and S. Kunte, "Understanding correlation," *Teach. Stat.*, vol. 24, no. 2, pp. 66–67, 2002.

[43] Wikipedia, " Cardiovascular drift", https://en.wikipedia.org/wiki/Cardiovascular_drift, January 08, 2019

[44] Wikipedia, " p-value", https://en.wikipedia.org/wiki/P-value, June 22, 2019

[45] Wikipedia, " Coefficient of determination", https://en.wikipedia.org/wiki/Coefficient_of_determination, June 12, 2019

[46] Wikipedia, " Convolutional neural network", https://en.wikipedia.org/wiki/Convolutional_neural_network, June 20, 2019

[47] Wikipedia, "Data preparation", https://en.wikipedia.org/wiki/Data_preparation, May 15, 2019

[48] I. Oksuz *et al.*, "Deep learning using K-space based data augmentation for automated cardiac MR motion artefact detection," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11070 LNCS, no. August, pp. 250–258, 2018.

[49] J. Hoffmann *et al.*, "Machine learning in a data-limited regime: Augmenting experiments with synthetic data uncovers order in crumpled sheets," *Sci. Adv.*, vol. 5, no. 4, 2019.

[50] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," pp. 1–15, 2014.

[51] Association for the Medical Advancement of Instrumentation (AAMI), *Diagnostic electrocardiographic devices, EC11 Diagnostic electrocardiographic devices*. 2001.

[52] B. Sjodin and J. Svedenhag, "Applied Physiology of Marathon Running," *Sport. Med. An Int. J. Appl. Med. Sci. Sport Exerc.*, vol. 2, no. 2, pp. 83–99, 1985.

[53] Wikipedia, "Anaerobic exercise", https://en.wikipedia.org/wiki/Anaerobic_exercise, June 10, 2019

[54] T. Soligard *et al.*, "How much is too much? (Part 1) International Olympic Committee consensus statement on load in sport and risk of injury," *Br. J. Sports Med.*, vol. 50, no. 17, pp. 1030–1041,

2016.

[55] T. J. Gabbett, “The training-injury prevention paradox: should athletes be training smarter and harder?,” *Br. J. Sports Med.*, vol. 50, no. 5, pp. 273–80, 2016.

[56] J. Margarito, R. Helaoui, A. M. Bianchi, F. Sartor, and A. G. Bonomi, “User-independent recognition of sports activities from a single wrist-worn accelerometer: A template-matching-based approach,” *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 788–796, 2016.