Towards a Warning System for Inaccuracies in PPG-based Heart Rate Measurements on Smartwatches: A Supervised Deep Learning Approach

Masterthesis

In Partial Fulfilment of the Requirements for the Degree of
Master of Science

University of Salzburg

Submitted by
Marlene Brunner, BSc

Supervisors:
Univ.-Prof. Dipl.-Math. Dr. Arne Bathke
Dr.-Ing. Jan David Smeddinck, BSc, MSc
Dr. Devender Kumar

Department of Artificial Intelligence and Human Interfaces

Salzburg, September 14, 2023

Abstract

Wearable devices that measure Heart Rate (HR) using Photoplethysmography (PPG), such as smartwatches or fitness bands, have become increasingly popular in recent years. However, the accuracy of PPG-based HR measurements can be affected by a number of factors. This thesis investigates the accuracy and validity of PPG-based HR measurements in comparison to gold standard Elektrocardiogram (ECG) readings. First background literature is summarized to investigate the current state of the art in PPG-based HR measurement. Then, a real world data set is analyzed to assess the accuracy of PPG-based HR measurements in a real-world setting.

The first part of the thesis shows that PPG-based HR measurements can be inaccurate, particularly at high exercise intensities. To address this issue, a Linear Regression model and two Deep Learning models have been developed to predict the measurement errors that occur based on the data stream from the wearable alone. The results show that a Deep Learning model based on a Convolutional Neural Network (CNN) outperforms the other models. The model is able to reliably detect large measurement errors and therefore it is possible to develop a warning system to inform the end user of the wearable when those large deviations from the gold standard occur.

Keywords: wearable devices, photoplethysmography, PPG, heart rate, HR, electrocardiogram, ECG, prediction, accuracy, validity, deep learning, CNN, transformer, warning system, machine learning, artificial intelligence, human-data interaction, digital health, smartwatch, fitness band

Acknowledgements

First and foremost, I would like to thank Univ.-Prof. Dipl.-Math. Dr. Arne Bathke, Dr.-Ing. Jan Smeddinck, BSc, MSc, and Dr. Devender Kumar for all their guidance, support, and excellent feedback. Their expertise and encouragement helped me to complete this research and write this thesis.

I would like to thank Univ.-Prof. Dr. Dr. Josef Niebauer, MBA, Priv.-Doz. Dr. Dr. med. Mahdi Sareban, Dr. Stefan Tino Kulnik, MRes, Michael Neudorfer, MSc, MEd, Bernhard Reich, PhD and Priv.-Doz. Dr. Gunnar Treff, for the opportunity to conduct my research at the Ludwig Boltzmann Institute and for the many resources and support they have offered me.

Finally, I would like to thank my entire family and friends for their constant support during my studies. Thank you for always having an open ear for me and giving me the strength and encouragement in times of doubting and thank you for always being there for me.

Contents

1 Introduction **1**

2 Theoretical Background **4**
2.1 Heart Rate 4
2.1.1 The Regulation of Heart Rate 4
2.1.2 The Measurement of Heart Rate 5
2.2 Time Series 8
2.3 Linear Regression 9
2.4 Artificial Neural Networks 12
2.4.1 Convolutional Neural Networks 17
2.4.2 Transformer 20
2.5 Human-Data Interaction 21

3 Evaluating the Accuracy and Validity of the Wearable **23**
3.1 Literature Research 23
3.2 Investigation of real-world data 24
3.2.1 Methods 24
3.2.2 Data Preprocessing 25
3.2.3 Descriptive Analysis 25
3.2.4 Results of Descriptive Analysis 27
3.3 Results 34

4 Prediction of Measurement Error **35**
4.1 Literature Research: Time Series Prediction 35
4.2 Methods 39
4.2.1 Models 39
4.2.2 Data Preprocessing 40
4.2.3 Data Splitting 40
4.3 Model Training 41
4.3.1 Multiple Linear Regression 41
4.3.2 CNN 42
4.3.3 CNN and Transformer 43
4.4 Results 45
4.4.1 Development of a Warning System 53
4.4.2 Evaluation of the Warning System 56

5 Discussion 58

6 Future work 66

7 Conclusion 67

List of Abbreviations IV

Bibliography VI

Appendix XXIII

1 Introduction

Monitoring HR using PPG has attracted much attention with the advent of wearable devices such as smartwatches and smartbands. Previously, HR had to be measured using ECG sensors that were attached to the chest and required ground and reference sensors [137]. In recent years, HR measurement using the PPG method has taken off. It is less expensive, easier to use, and does not require ground and reference sensors [93]. PPG is easily integrated into smartwatches and wristbands, providing a non-invasive and indirect estimation of HR [9]. These so-called wearables can be unobtrusively attached to various body sites, such as an earlobe, fingertip, or wrist.

Wearables have gained popularity in both hobby and professional sports as well as in the health and medicine industry. Health professionals enjoy the ability to use them to monitor patients individual internal response during recovery, post-op, sleep, or even medication intake [112]. The use of wearables has been increasing in recent years and was the top fitness trend in 2022 [154]. In this context, Patel et al. [116] investigates whether a person's health behavior changes as a result of using wearable devices. The study found that in some cases it does, and wearable devices can effectively promote health behaviors. Wearable devices can help people increase their physical activity and can also be a helpful way to track individual progress to stay motivated. Furthermore, wearables can provide feedback that helps people identify areas where they can make changes. Thus, wearable devices can be a significant way to improve health by increasing physical activity or by otherwise accompanying physical activity.

The benefits of regular physical activity are well known and well documented. Physical activity is an important modifiable risk factor for a number of chronic diseases (cardiovascular, stroke, type 2 diabetes, cancer, obesity) and all-cause mortality. Furthermore, physical activity can improve mental health and quality of life. [162]

Physical inactivity, in turn, is the leading cause of Cardiovascular Disease (CVD). CVDs are a group of diseases that affect the heart and blood vessels, affecting it so that it cannot perform normal functions. CVDs are the leading cause of death worldwide. An estimated 17.9 million people die every year from CVD, accounting for 32% of all deaths worldwide. Early diagnosis of CVD is essential to improve patient outcomes. HR is an important indicator of cardiovascular health as having a high

HR can increase the risk of heart attack, stroke and CVD. Therefore, it is important to monitor the HR to detect irregularities at an early stage in order to change the lifestyle accordingly. Early diagnosis is very significant for treatment, as it is most effective in the early stages of the disease. By monitoring HR, wearables can help detect early signs of CVD. [5, 39]

In addition to monitoring HR for prophylaxis or early detection of disease, HR is commonly used for monitoring cardiovascular exercise intensity. However, previous studies have shown that HR measurements can be inaccurate, particularly at high exercise intensities. This is because HR is sensitive to body movements and other factors, such as skin temperature and hydration status. [33, 57, 95, 102]. The inaccuracy of HR measurements can have implications for training prescription. For example, if a person's HR is measured incorrectly, they may be prescribed an exercise intensity that is either too high or too low. This could lead to injuries or suboptimal training results [160].

Studies and analyses are available that investigate the accuracy between the ECG as gold standard and PPG measurements [27, 58, 68, 102, 153, 160]. The growing interest in wearables and the ever-increasing research interest in them is bringing rapid growth in data. This provides an opportunity to apply Machine Learning or Deep Learning techniques to extract insights from such datasets.

To date, frequently occurring divergences in PPG-based HR measures from wearables are known, but there is no way to report this divergence back to the user. There is only a very limited amount of work that actually perform predictions for continuous values of physiological data, such as HR [97]. However, recent advances in Artificial Intelligence (AI) are revolutionizing the healthcare industry. One of the most promising applications of AI in healthcare is to improve the utility and reliablity of wearable sensors. The use of AI and wearable sensors are still in their infancy, however by providing reliably predicted offsets, they have the potential to revolutionize the healthcare industry by using this predicted offsets to implement a practical warning system and communicate those to the end user.

The goal of this thesis is to not only investigate the validity and verify the accuracy of the wearable, but also to use Deep Learning techniques to estimate the current measurement error and provide this information to the user. Not only the detection of a measurement error but also the communication to the end user is an important point here since this human-data interaction is of growing relevance in the field of digital health. [25]

The thesis is divided as several sections: Chapter 2 gives the theoretical background on the topics that play an important role in this thesis. Chapter 3 deals with the investigation of the validity of wearables, where first a literature research was conducted and then a real data set was evaluated. An important part of this thesis was the development of different models to predict HR data, which is explained in Chapter 4. In this chapter, first the latest findings in the literature are shown, then the applied models are explained in more detail including the description of the training of the models and finally the results are presented. Finally, a discussion of the results found as well as possibilities for further research complete the thesis. This can bridge the gap that currently exists between the practical application of HR measurements using PPG and the associated measurement errors and reduced validity as a feedback to the end user.

2 Theoretical Background

In the following chapter, the theoretical background of HR, time series, Artificial Neural Network (ANN), especially CNN and Transformer, Linear Regression and Human-Data-Interaction are briefly presented, since they play an important role in the context of this master thesis. The state-of-the-art literature follows later in Chapter 3.1 and 4.1.

2.1 Heart Rate

All cells in the body, in order to be able to guarantee the maintenance of metabolism and thus the maintenance of life, must be supplied with oxygen. This happens through the transport of oxygen-enriched blood, which is pumped through the body by the heart. The HR indicates how often the heart chambers contract and relax within one minute [112].

2.1.1 The Regulation of Heart Rate

HR is tightly controlled by several biological systems through various feedback loops that ultimately regulate and maintain energy homeostasis [2]. Without these control mechanisms, the HR would be limited to an intrinsic value of 100 to 110 Beats per Minute (bpm), therefore, adapting poorly to changing energy demands [69]. The heartbeat regulation is assisted by the Autonomic Nervous System (ANS), which attempts to minimize the energy expended to ensure a heartbeat, while also being able to meet the immediate needs from the external environment [109]. The ANS is stimulated, in part, by perceived stressors in the external environment. These include physical stressors (increased oxygen demand due to increased metabolism exempli gratia (e.g.) through exercise), environmental stressors (temperature changes, altitude changes, noise, et cetera (etc)) or psychological stressors (anxiety, fear, etc) [22, 135]. In addition, a wide variety of internal physiological conditions such as fluid intake, nutrient availability, hormonal sensitivity and fatigue can alter the sympathetic response [30, 61, 83, 156].

The ANS regulates the heartbeat via the autonomic nerves and its two branches: the sympathetic and parasympathetic nervous systems [130]. While the sympathetic branch initiates the release of a cascade of hormones, such as norepinephrine, which leads to an increase in HR, the parasympathetic branch plays the main role. It restores homeostasis after periods of stress and provides energy for cell repair, regeneration and adaptation [51, p. 177-78][155]. Together, the sympathetic and parasympathetic systems drive the body's adaptive mechanisms that allow it to be better equipped to cope and perform in different environments. The relationship between the two branches and their responses to acute and chronic stress play a critical role in overall health, performance and risk of injury [146, 166]. Thus, monitoring changes in HR in different contexts is the most common and proven method to observe this balance. Insights can be gained into health and even athletic performance with the goal of optimizing it [22, 118, 120, 159]. Heart rate can be viewed as the sum of the body's responses to physical and mental stress [131].

2.1.2 The Measurement of Heart Rate

The use of HR measurements began with the introduction of the first wireless ECG chest strap in 1983 by Polar Electro [52]. Since then, the need for HR measurements has become more prevalent. The advent of a variety of mobile, low-cost applications has led to accelerated and improved research and development of these measurement devices [50, 164]. Monitoring HR has several advantages, including ease of recording, noninvasive nature and cost efficient. In addition, it can be measured over multiple time periods and physiological conditions [51, p. 178].

With the proliferation of wearable sensors, healthcare and clinical examination procedures have seen further improvements [59]. Health monitoring systems can monitor a patient's cardiovascular status at home and provide recommendations to both the patient and healthcare provider [70]. Rapid technological advances in this field have now brought vast amounts of health-related data. These data play an important role in early and accurate detection and diagnosis of diseases for personalized treatment and prognosis assessment [112]. For example, according to the World Health Organization (WHO) [39], high HR increases the risk of death, heart disease and cardiovascular disease. Monitoring HR is therefore essential to detect irregularities early in HR in order to counteract health problems at an early stage. The earlier heart disease can be detected and treated, the better, since treatments are most effective in the early stages [5].

Measuring HR can be done using either an ECG or a PPG sensor. In 1895, Dr. Willhelm Einhoven measured heart activity for the first time using an electrical signal and was awarded the Nobel Prize in 1924. [51, p. 178]. Today, the ECG is one of the most powerful diagnostic tools in modern medicine. [72].

An ECG consists of a linear recording of electrical activity of the heart. For each cardiac cycle, an atrial depolarization wave (P wave), a ventricular depolarization wave (QRS complex), and a ventricular repolarization wave (T wave) are recorded. In a normal rhythm, the sequence is always P-QRS-T, which can be seen in Figure 2.1. The intervals between the waves in a cycle are variable, depending on HR and rhythm. The theory behind an ECG recording is that the ECG is an expression of the electro-ionic changes that occur during depolarization and repolarization of the heart muscle. In both phases, electrical charges and currents are formed in the heart muscle and can be measured with the help of sensors. [35].

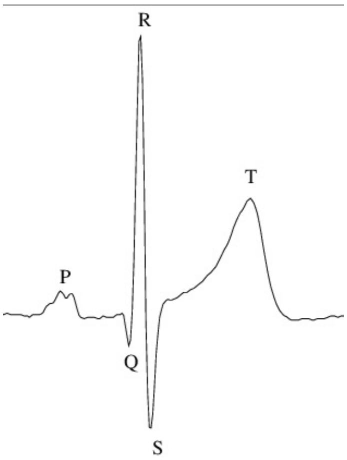


Figure 2.1: The P-QRS-T complex of a normal ECG wave. Adapted from Zhang ([174]

In contrast, PPG detects blood volume changes in the tissue's microvascular bed [28]. The basic form includes two components: a light source to illuminate the tissue and a photodetector to measure the small variations in the intensity of the light associated with the changes in blood flow. The main factors that can affect the amount of light received by the photodetector are blood volume, blood vessel wall motion and red blood cell orientation [6]. PPG signals are optically detected by pulse oximeters. The pulse oximeter illuminates the wearer's skin with a Light-Emitting Diode (LED). As a result of reseach conducted over the last decades, primarily green light is used, which has a shorter wavelength compared to red or infrared. Thus, large intensity variations can be generated in cardiac modulation and it provides better Signal-to-Noise Ratio (SNR) [94, 176]. The photodetector then measures the intensity changes of the light reflected from the skin. This produces a PPG signal [6, 71]. In addition, a reflective system, in which the LED and photodetector are on the same side, is preferable because it provides more comfort to the user [148]. The periodicity of the PPG signal corresponds to the heart rhythm, so HR can be estimated from this signal (Fig. 2.2) [176]. The pulse can be divided into anacrotic and catacrotic phases. The anacrotic phase is the rising part of the signal, and the catacrotic phase is the falling part. The first phase mainly involves the systole[1] of the heartbeat, the second phase involves the diastole[2] as well as the wave reflections from the periphery [6].

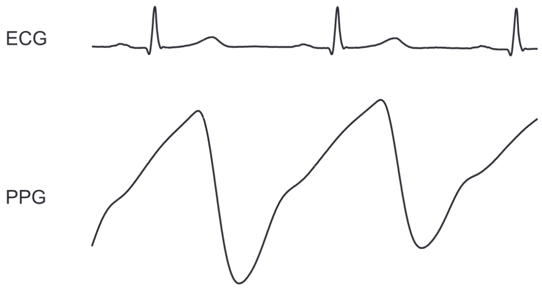


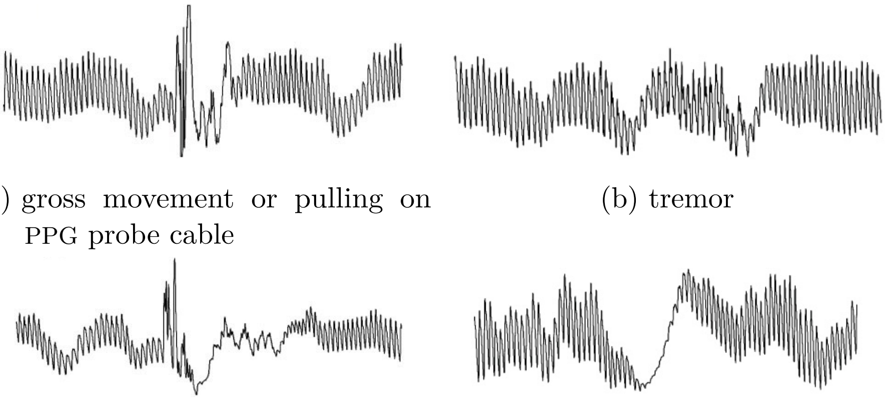
Figure 2.2: The PPG signal compared to the corresponding ECG recording of the HR. Adapted from Allen et al. [6].

In addition to measuring HR, PPG technology allows the measurement of HR variability, oxygen saturation, blood pressure, cardiac output, assessment of autonomic function, and also for detection of peripheral vascular diseases. This success is possible, even though the properties of the PPG waveform are not fully understood [6].

PPG signals can be recorded inconspicuously at various body sites, e.g., the earlobe, fingertip, or wrist [5, 9]. However, due to the slight distance between the sensor and the skin surface, the measurements can be contaminated by movements, called Motion Artifact (MA). These contaminations can also be caused by abnormal blood pressure changes, which makes accurate HR estimation very difficult, especially during intense exercise [5, 7, 46, 176]. The wrist can cause much stronger and complicated motion artifacts, compared to fingertips and earlobes, due to higher flexibility. However, this position facilitates the design of wearable devices and maximizes usability; therefore, the development of powerful algorithms for monitoring and validation for PPG signals are of great value [176]. Figure 2.3 shows examples for different types of MA caused by different reasons.

[1] During systole, the heart contracts, thereby pumping blood into the systemic circulation.

[2] In diastole, the heart relaxes and blood can flow from the veins back into the heart.



(a) gross movement or pulling on PPG probe cable

(b) tremor

(c) coughing

(d) changes in breathing pattern (e.g. yawning)

Figure 2.3: Examples of different types of MA on index finger. Adapted from Allen et al. [6]

2.2 Time Series

A time series is an ordered sequence of real-value variables and represents a collection of chronological observations. Values are collected from measurements made at uniformly distributed time points at a given rate. A time series is thus a set of contiguous time points which can be either univariate or multivariate. Shumway et al. [138] define time series as "a collection of random variables indexed according to the temporal order of their occurrence".

A univariate time series

$$\mathrm{X} = [x_1, x_2, ..., x_T]$$

is an ordered set of real values. The length of X is equal to the number of real values T.

An m-dimensional multivariate time series

$$\mathrm{X} = [X_1, X_2, ..., X_M]$$

consists of M distinct univariate time series with $X_i \in \mathbb{R}_T$ [49].

A time series represents a collection of chronological observations. Big data size, high dimensionality and continuous updating are characteristics of time series data.

These data are always considered as a whole rather than as individual numerical fields because of their numerical and continuous nature [53].

2.3 Linear Regression

A basic method for predicting the future from past data is Linear Regression. Linear Regression is a statistical technique that examines and models the relationship between variables. It is used in many fields, including engineering, physical and chemical sciences and even life and biological sciences. It is perhaps the most widely used statistical technique. Regression is used for data description, parameter estimation, prediction and estimation, and control, among other applications. [127]

The following statements are based primarily on Montgomery et al. [134] and Rencher et al. [127].

Linear Regression is distinguished between different models. In all models, X is referred to as the independent variable, predictor, or regressor variable, and y is referred to as the dependent variable or response variable.

When only one regressor variable is included, it is referred to as a **Simple Linear Regression**.

$$y = \beta_0 + \beta_1 x + \epsilon \tag{2.1}$$

The equation 2.1 describes a linear regression model, where the intercept β_0 and slope β_1 are unknown constants and are called regression coefficients. The slope is the change in the mean of the distribution of y caused by a one-unit change in x. ϵ describes the error term of the model, which represents random fluctuations, measurement errors or the effect of external factors.

For n observations, the Simple Linear Regression model according to 2.1 can be written as follows:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, 2, ..., n \tag{2.2}$$

In this context, 'simple' means that there is only one prediction variable and 'linear' means that the model is linear in β_0 and β_1. In addition, the following assumptions are made:

1. $\mathbb{E}(\epsilon_i) = 0$ for all $i = 1, 2, ..., n$, or, equivelantly, $\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i$.

2. $var(\epsilon_i) = \sigma^2$ for all $i = 1, 2, ..., n$, or, equivelantly, $var(y_i) = \sigma^2$.

3. $cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$, or, equivelantly, $cov(y_i, y_j) = 0$.

Assumption 1 states that y_i depends only on x_i and all other variations in y_i are random. Assumption 2 states that the variance of ϵ or y does not depend on the values of x_i (=homoscedasticity). Assumption 3 states that ϵ or the y variables are not correlated with each other.

The estimated model can then be used to draw conclusions such as confidence intervals or hypothesis tests or to predict the value y for new values x.

The least squares method is used to estimate β_0 and β_1. This involves looking for values that minimize the sum of the squares of the deviations $(y_i - \hat{y}_i)$ of the n observed values (y_i) from their predicted values $(\hat{y}_i = \beta_0 + \beta_1 x_1)$.

$$\hat{\epsilon}'\hat{\epsilon} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta_1 x_i})^2 \tag{2.3}$$

When a regression model contains more than one regressor variable, it is called a **Multiple Linear Regression (MLR)**.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon \tag{2.4}$$

The model 2.3 is a Multiple Linear Regression model with k regressors, where $\beta_j, j = 0, ..., k$ are the regression coefficients. This model describes a hyperplane in a k-dimensional space. The parameter β_j describes the expected change in y per unit change in x_j when all other regression variables $x_i, i \neq j$ remain constant.

In the Multiple Linear Regression model, the assumptions for ϵ and y_i remain essentially the same as for Simple Linear Regression:

1. $\mathbb{E}(\epsilon_i) = 0$ for all $i = 1, 2, ..., n$, or, equivelantly, $\mathbb{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$.

2. $var(\epsilon_i) = \sigma^2$ for all $i = 1, 2, ..., n$, or, equivelantly, $var(y_i) = \sigma^2$.

3. $cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$, or, equivelantly, $cov(y_i, y_j) = 0$.

For n observations, the above formula is as follows:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_k x_{1k} + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_k x_{2k} + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + ... + \beta_k x_{nk} + \epsilon_n \end{aligned} \tag{2.5}$$

These n formulas can also be written as a matrix:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1k} \\ 1 & x_{21} & x_{22} & ... & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & ... & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{2.6}$$

or as

$$y = X\beta + \epsilon \tag{2.7}$$

Both Simple Linear Regression and Multiple Linear Regression take the basic assumption that the relationship between variables is linear. However, if the functional relationship of the variables x and y is not linear, then both models have difficulty providing good estimates and models. In this case, **Polynomial Regression (PR)** can be used since complex nonlinear relationships can be well modeled by polynomials over relatively small ranges of x-values.

Polynomial Regression models can contain either one variable (2.8) or multiple variables (2.9). Here k describes the order of the polynomial regression.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^k + \epsilon \tag{2.8}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_2 x_2^k + \beta_{12} x_1 x_k + \epsilon \tag{2.9}$$

2.4 Artificial Neural Networks

The following statements are based primarily on the work of Nielsen [107], Bishop [17], Walczak et al. [158] and Kruse et al. [82]. Artificial Neural Networks (ANNs) are systems that record, process and transmit information and whose structure is modeled on the nervous systems of animals and humans. They consist of relatively simple units called neurons that operate in parallel to produce one common output. Neurons communicate via connections in the form of activation signals.

An artificial neuron, also called a perceptron, generates a single binary output from n binary inputs, $x_1, ..., x_n$:

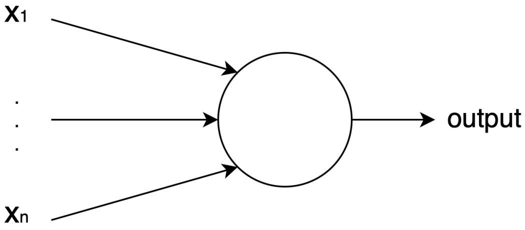


Figure 2.4: Sketch of a Perceptron

With the help of n weights $w_1, ., w_n$, the importance of the respective input for the output can be expressed. The vectors w and x are vectors describe the weights and inputs, which are multiplied and summed up to get a weighted sum. The output of the neuron is defined with the help of a threshold value. Depending on whether the weighted sum $\sum w_j x_j$ is greater or less than the threshold, 0 or 1 is the output:

$$output = \begin{cases} 0, & \text{if } \sum_j w_j x_j \leq \text{threshold} \\ 1, & \text{if } \sum_i w_i x_i > \text{threshold} \end{cases} \tag{2.10}$$

The weights as well as the threshold are real numbers describing the parameters of the neuron. If now $w_j x_j$ is changed to $w * j = w_j x_j$ and the threshold is brought to the other side of the inequality, it can be replaced by the so-called bias of the perceptron $b \equiv -\text{threshold}$. The perceptron rule 2.10 thus changes to:

$$output = \begin{cases} 0, & \text{if } w * x + b \leq 0 \\ 1, & \text{if } w * x + b > 0 \end{cases} \tag{2.11}$$

Here, the bias describes the measure of how easily the perceptron outputs a 1.

In addition to perceptrons, there are sigmoid neurons, which are modified so that small changes in their weights and biases also cause small changes in their output.

This is critical for a network to learn. Like a perceptron, the sigmoid neuron can have one or more binary inputs and produces a single output. The difference here is that a perceptron can only produce values of 0 or 1, while sigmoidal neurons can also produce values in between.

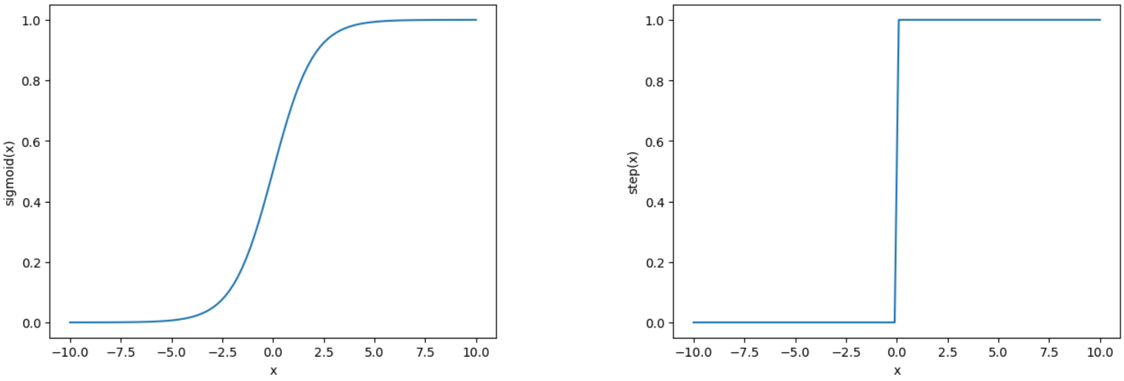
The sigmoidal function ($\sigma(\cdot)$ is defined by

$$\sigma(z) \equiv \frac{1}{1+e^{-z}} \tag{2.12}$$

The output of a sigmoidal neuron with n inputs $x_1, ..., x_n$, n weights $w_1, ., w_n$ and a bias b is thus:

$$\text{output} = \frac{1}{1 + exp(-\sum_j w_j x_j - b)} \tag{2.13}$$

Figure 2.5a shows the form of a sigmoidal function while the function of a perceptron is a simple step function, which can be seen in Figure 2.5b.



(a) Sigmoid Function

(b) Step Function

Figure 2.5: Comparison of the function of a sigmoid neuron (left) versus the function of a perceptron (right)

Each neuron has three functions: the network input function, the activation function and the output function. The operation of an ANN, simply stated, works as follows: each neuron receives an output from the previous neuron with the associated weights. From this, the neuron calculates the network input with the help of the input function. The activation function calculates the new activation of the neuron, from which the output function calculates the output of the neuron.

The computations of a neural network can be divided into two phases: the input phase and the work phase. The input phase in where the external inputs are fed into the network, and the work phase is where the output is computed. In the input phase, the activations of the input neurons are set to the values of the corresponding external inputs and the output function of these neurons is produced. In the work phase, the activations and outputs of each neuron are computed (input function, activation function, output function). As soon as the recalculations have reached a stable state or a predefined number of recalculations has been performed, the recalculations are terminated.

The number of layers of processing elements or nodes, including input, output and possible hidden layers, as well as the number of nodes contained in each layer determine the architecture of the ANN. Figure 2.6 shows a simple representation of an ANN.

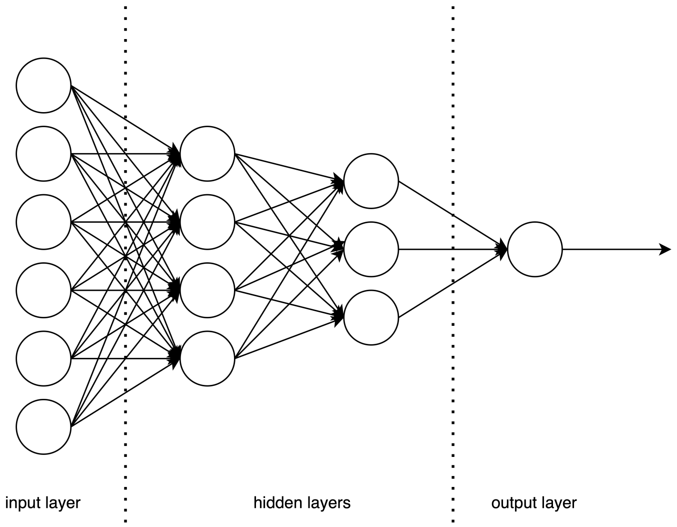


Figure 2.6: Basic architecture of a Artificial Neural Network

The neural network is composed of three types of neurons: input neurons, output neurons and hidden neurons. Input neurons receive data from the environment, output neurons send data back to the environment and hidden neurons lie between the input and output layers and perform intermediate computations. The hidden neurons are not directly connected to the environment, hence the name "hidden".

ANNs can be built without hidden layers, but in application they are diminished accordingly, as is this case, where they can only classify input data that is linearly separable. In order to solve nonlinear and complex problems, multiple hidden layers can be used. The number of hidden layers is related to the complexity of the problem to be solved. More layers can increase the accuracy of the fit, while a smaller number of layers improves the extrapolation capabilities. Here, the number of layers is determined heuristically. As the dimensionality of the problem space increases - higher order problems - the number of hidden layers should increase accordingly. If the number of nodes for a hidden layer is determined, the number of hidden nodes increases the training time. At the same time more feature detectors can be used, however, too many can lead to overfitting and thus poor generalization performance.

Depending on the network structure, two types of ANN can be distinguished:

- Feed-forward network: In the case, there are no cycles or loops, where a loop is a connection from a neuron to itself.
- Recurrent network: Loops or directed cycles occur.

If the network structure is acyclic, the direction of information transfer is exclusively from input neurons to output neurons. However, if there are loops or directed cycles, the outputs can be recurrent with the inputs.

Two different types of learning can be distinguished, depending on the training data and the criterion to be optimized: free learning and fixed learning. For a fixed learning task, the neural network is trained to produce a corresponding output in the output vector for every external input in the input vector for all training patterns. Since in practice this optimum can rarely be achieved, an error function is used that compares the desired outputs with the actual outputs to measure the match. This error function is usually defined as the sum of squared deviations between the desired and actual outputs over all training patterns and all output neurons. Thus, a fixed learning task has a desired output and allows to compute errors.

In contrast, a free learning task for a neural network with n input neurons is a set of training patterns, where each pattern consists of an input vector. Here, a different criterion is needed to assess how well the ANN can solve the task. For example, an example of this learning task is the creation of clusters of similar vectors (clustering). Here, the idea is to generate "similar outputs for similar inputs". The similarity between the outputs of a cluster should be as small as possible, while the similarity between the different clusters should be as large as possible. This can be defined with the help of a distance function.

The basic principle of training an ANN is to adjust the connection weights and other parameters (e.g. thresholds) in order to optimize a certain criterion. To improve the performance of ANN, the network must train and learn. Training consists of changing the weights in the network to the point where the best result can be achieved. After each output, it is compared to the desired output and a total error is calculated. The smaller the value of the total error, the better the network. Back propagation is used to change the weights and thresholds so that the error gradually becomes smaller. If the value no longer changes, the training process is complete. In this process, back-propagation uses gradient descent as an optimization technique. Figuratively, this can be seen as a search for the global minimum on an error surface. Often, the quadratic cost function C is used as error function for this purpose, which has the following form:

$$C = \frac{1}{2n} \sum_x ||y(x) - y'(x)||^2 \tag{2.14}$$

C describes the average loss over n training examples where $y(x)$ describes the output of the carry function from $y = f(\sum w * x + b)$ and $y'(x)$ represents the desired output. By calculating the partial derivative of C with respect to each weight (and each bias), the total error can now be minimized, as shown in equation 2.15

$$\frac{\partial C}{\partial y} = y - y' \tag{2.15}$$

With the calculated gradient of each training sample it can now be determined if and how the weights and the bias should be changed. Thus global minumum can be found.

Design parameters for ANNs

In order to find a suitable model for a prediction, design parameters have to be defined. The selection of these are essential for success, as unsuitable parameters may result in the network being unable to train. Besides the number of layers and neurons and the selection of the activation function, techniques for regularization, the size of the learning rate, the epochs and the stack size can be defined.

In order not to fit the model too much to the training data and thus get an overfitting, there is the possibility of regularization. The most common technique used is the "dropout technique", which has the idea of removing random neurons with their connections from the network during the training phase. This makes the network more robust and insensitive to weights of the other neurons.

The learning rate controls the adjustment of weights and bias with respect to the loss gradient. It determines the size of steps the model takes toward the local minimum. If the learning rate is too small, the optimization takes a very long time and the model runs the risk of getting stuck in the local minimum. However, if the steps are too large, the model may miss the global minimum and the loss may even increase.

The number of epochs describes the number of times the entire data set has been run through the model. After each epoch, the weights and biases can be adjusted and optimized. More epochs also mean more opportunities to find better values for weight and bias, but this also increases the training time and runtime. The size of a stack describes the number of training samples that are used within an epoch. The iterations describe the number of times a stack is run through the ANN.

2.4.1 Convolutional Neural Networks

Since their introduction in the 1990s, Convolutional Neural Networks have contributed enormously to the success of machine learning. CNNs are designed to mimic the way the human brain thinks. They learn fully automatically, which allows them to extract features that are salient in the input data across different layers. [123]

Convolutional Neural Networks are a special type of ANNs that have a grid-like topology. Here, time series data is represented as a 1D grid, while image data is represented as a 2D grid of pixels. A CNN uses a mathematical operation called 'convolution' in at least one of its layers. This is a special kind of linear operation.

Convolution is an operation on two functions with a real argument. The goal here is to average several measurements. By weighting the measurements, they can be included in the calculation depending on their relevance. This can be achieved with a weighting function $w(a) : s(t) = \int (x(a)w(t-a)\partial a$. This operation is called convolution and is typically notated with an asterisk: $s(t) = (x * w)(t)$. The first argument (here the function x) is called the input, and the second argument (here w) is called the kernel. The output is sometimes referred to as the feature map.

The advantages of convolution are the following three important ideas:

- Sparse interaction:

  Traditional ANNs operate with matrix multiplication in which each output unit interacts with each input unit. CNNs operate with a sparse interaction in that the kernel is smaller than the input. As a result, fewer parameters are stored, which improves both memory requirements and statistical efficiency by requiring fewer operations to compute the output. With m inputs and n outputs, matrix multiplication requires $m \times n$ parameters and thus has a running time of $O(m \times n)$. If the number of connections is limited to $k(k < m)$, as is the case with convolution, the parameters decrease to $k \times n$ and the running time decreases to $O(k \times n)$.

- Parameter sharing:

  Whereas in a conventional ANN each element of the weight matrix is used exactly once to compute the output of the layer, a CNN uses each part of the kernel at each position of the input. This eliminates the need to learn a separate set of parameters for each position. This further reduces the memory requirement.

- Equivariant representations:

  The parameters are split in a special form in the case of a convolution, so that the shift can be called equivariant to the translation.

  Definition: A function $f(x)$ is equivariant to a function g if $f(g(x)) = g(f(x))$.

  In the case of convolution, the function is equivariant to g if g is any function that translates, id est (i.e.), shifts, the input. In the case of time series data, this means that convolution produces a kind of timeline representing the time of occurrence of different features in the input. If an event is shifted back in time in the input, it will appear in the output in exactly the same representation, only later in time.

In addition, convolution can work with variable-size inputs.

The architecture of the network consists of several layers. The basic architecture of CNNs is single-headed. However, adding multiple heads multiplies pattern learning. Each head can have different filter banks and different processing layers. In addition, a pooling or dropout layer can be added in each head. These combination options can achieve better performance and hence better learning results [8, 169].

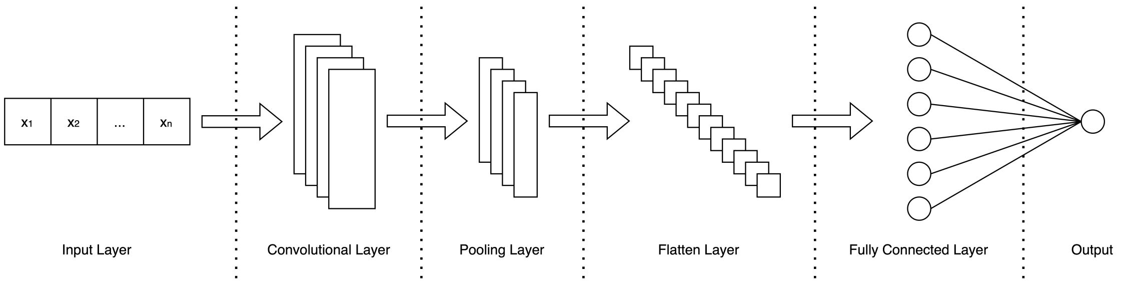


Figure 2.7: Architecture of a simple CNN

Figure 2.7 shows the architecture of a simple CNN. The first layer of the CNN is the input layer, which receives the raw time series data. The next layer is the convolutional layer, which applies a series of filters (kernels) to the input data. Each layer consists of three stages. In the first stage, multiple convolutions are performed in parallel to generate presynaptic activations. For time series data, each filter slides over the time series, performs element-by-element multiplication, and sums the results to produce a single output value. By applying multiple filters to the input data, the convolution layer produces a series of output feature maps that highlight different temporal patterns in the time series. In the second stage, each activation is guided by a nonlinear activation function. The intent behind this is to be able to adapt to a wider range of activation functions, rather than being limited to linear functions. Recitifed Linear Units (ReLUs) are often used here.

$$\text{ReLu} = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases} \qquad = max(0, x)$$

In the third and last stage, a pooling function is applied. Here, an input vector or matrix is divided into several regions, where each region contains several values and are combined into a single value by the pooling operation. The most common variants of pooling are max-pooling or mean-pooling, by using the maximum value or the mean value with respect to each region. Other pooling operations would still be the L2 norm or a weighted average. In all cases, pooling helps to make the representation invariant to smaller shifts in the input. The idea is that relevant information outweighs irrelevant information and continues to be included in the output. Multiple pooling stages further reduce the impact of small changes due to translation, rotation or scaling. After one or more of these convolutional layers, a pooling layer is added to reduce the dimensions of the feature map while preserving the most relevant information. Next, the flatten layer takes these outputs and combines all the extracted features into a single vector. This vector is then fed into one or more Fully Connected Layers (also called Dense Layers) to learn higher level representations by connecting all the extracted features. The final Fully Connected Layer is usually followed by a softmax activation function for classification tasks or a linear activation function for regression tasks. Finally, the output layer produces the final predictions based on the learned representations from the previous layers. [14, p. 306]

Figure 2.8 shows the combination of the individual layers, where each convolution layer consists of the three previously defined stages.

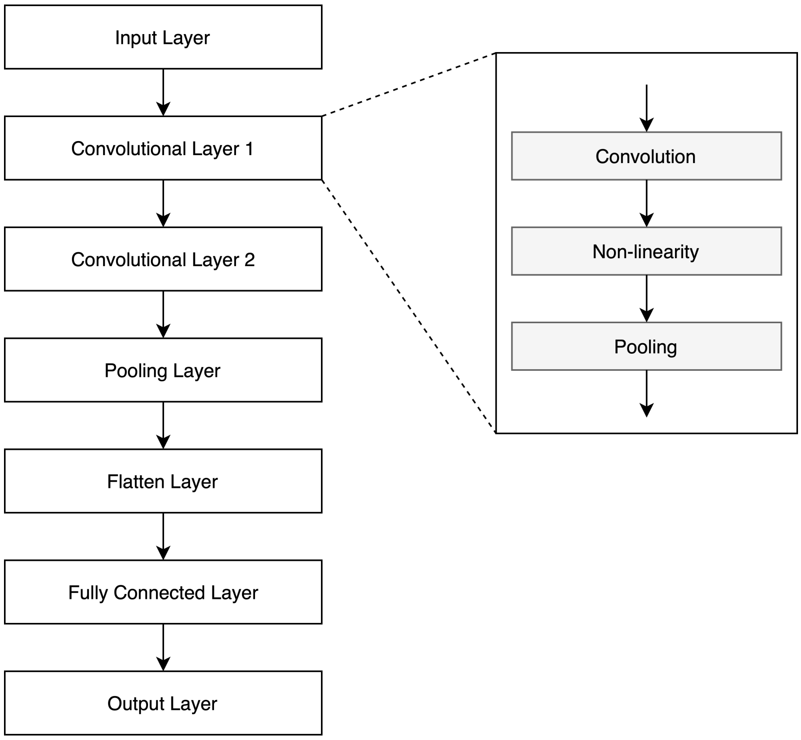


Figure 2.8: Each Convolution Layer involves Convolution, Non-linearity and Pooling.

2.4.2 Transformer

The following statements are based primarily on the work of Bahdanau et al. [12], Chandra et al. [29], Cheng et al. [32], Katrompas et al. [73], Kim et al. [78], Li et al. [88] and Zhao et al. [178]. The basic idea of a Transformer model is an encoder-decoder model, where the encoder represents the input from outside as an internal representation and transfers it via the decoder to the output sequence. The transformer model does not use layers but works with the architecture component called attention. This component allows interaction between each pair of tokens in a sequence allowing the model to learn relationships between tokens that are relevant to a task. In this process, most attention modules automatically learn a pattern of interaction between pairs of tokens. These patterns give each input token a weight for the specific prediction task. This enables the model to learn dependencies between tokens even if they are far apart in the input sequence. Modeling dependencies without regard to their distance in the input or output sequences is a major advantage of these transformers. They are potentially better at capturing recurrent patterns with long-term dependencies. Most often, several of the attention modules are used in parallel, allowing the model to learn different relationships. This is then referred to as multi-head attention. As a result, Transformer models are able to focus on different aspects of relevance between input elements, effectively capturing different representative subspaces.

2.5 Human-Data Interaction

The following statements are based primarily on the work of Kitchenham and Ensio-Lehtonen [80, p. 1148-1167] and Cabitza and Locoro [25],. Human-Data Interaction (HDI) is the study of how humans interact with data, both in terms of how they create and use data, and how they are affected by the data they interact with. In the health sector, HDI is concerned with how patients, healthcare providers, and other stake-holders interact with health data. HDI is a relatively new field of research, but it is growing rapidly due to the increasing amount of health data being generated and the growing importance of data-driven decision-making in healthcare. HDI research has the potential to improve the quality of care, reduce costs, and improve patient outcomes.

Some of the key principles of HDI in the health sector include:

- **Empowerment:** HDI should empower patients and healthcare providers to control their own data and to use it to make informed decisions about their health.

- **Transparency:** HDI should be transparent about how data is being collected, used, and shared.

- **Privacy:** HDI should protect the privacy of patients and other individuals whose data is being used.

- **Accuracy:** HDI should ensure that the data being used is accurate and reliable.

- **Usability:** HDI should design systems that are easy to use and understand.

In the health sector HDI is primarily used to develop personalized medicine approaches that thake into account the individual patient's genetic, environmental and livestyle factors. HDI is also used to develop clinical decision support systems that help healthcare providers making better decisions. It is used to engage patients in their own care, providing them access to their health data and tools to help them manage their health. Another approach is to deliver telehealth services, which allow patients to receive care from distance.

However HDI in the health sector still has some difficulites. Patients and healthcare experts have to be willing to use new HDI technologies, which is difficult if they are not familiar with them or if they do not trust them. There can occure concerns about privacy, security or the accuracy of the data. HDI systems also must be easy to use and understand. If systems are too complex or difficult to use, patients may not be able to understand them effectively, which can lead to errors and misunderstandings.

HDI systems also must provide a good user experience. They should be easy to navigate and responsive and they have to be user-friendly, otherwise patients will not be willing to use them. HDI systems must be designed to prevent misrepresentations, misinterpretations, and misunderstandings of health data. This can be done by using clear and concise language, providing context for the data, and allowing users to ask questions. What is more, there are no universally accepted standards for digital health data, which can make it difficult to share and integrate data from different sources. Different digital health systems may not be able to communicate with each other, which can make it difficult to get a complete picture of a patient's health. There are different regulations governing the use of health data in different countries, which can make it difficult to develop and deploy HDI systems.

3 Evaluating the Accuracy and Validity of the Wearable

This chapter aims to evaluate the validity of PPG measurement by first reviewing the literature and then analyzing real data.

3.1 Literature Research

Many different studies by different research groups showed unsatisfactory results of HR agreement of PPG measurements in comparison to ECG measurements [24, 27, 54, 68, 143].

Navalta et al. [105] investigated the validity of HR from various wearable devices with PPG measurement during a trail run with variable intensities. They found that the PPG devices, regardless of their position (finger, wrist, ear, forearm), did not provide acceptable agreement compared to the gold standard during runs of less than 20 minutes. They also observed that the validity during treadmill training was higher than during unrestricted activities. Especially in high intensity ranges, the accuracy decreases in their study. Bunn et al. [24] found in their study, that the best results were achieved at rest or when training on a bicycle ergometer. The higher the intensity, the lower the accuracy. Also, Thiebaud et al [153], Dondzila et al [40], Reddy et al [125], Düking et al [43], Jo et al [68] stated that intensity ranges matter most and the highest intensity range shows the highest measurement inaccuracies. Thiebaud et al. [153] compared PPG estimations with an ECG measurement while running on a treadmill at different speed and intensity ranges. They concluded that the accuracy of the devices may not yet be high enough to use it in research or to recommend it to athletes who need precise HR measurements for training purposes. Dondzila et al. [40] measured HR via PPG devices while walking and running. With increasing intensity, the accuracy of a wearable decreased. Reddy et al. [125] compared the PPG measurements over two days while performing different tests, training sessions and activities of daily living. They compared the measurements to those of a cheststrap. The wearables were reasonably accurate at measuring HR however there was more error observed when training in a high intensity range and with less wrist motion. Düking et al. [43] investigated the accuracy of four wrist-worn wearables while sittinig, walking and running at different intensities. They stated that the measurements of two of these wearables at high intensity range should be interpreted with caution since the error rate increased. Jo et al. [68] measured the HR via two wearables while resting, cycling, walking, jogging, running, arm raises, lunges and isometric plank. Both wearables showed a decrease in accuracy during higher intensities while one of them failed to satisfy their validity criteria.

Nevertheless, there are also some research groups that obtained satisfactory results in their studies regarding the measurement accuracy of PPG sensors [114, 145, 160]. However, the studies are designed differently, making it difficult to compare the results. The validity of PPG measurement can be affected by a number of factors, including the size and generalizability of the subject group, the length and performance of the measurement and examination, and the use of different reference measurements in those investigations.

3.2 Investigation of real-world data

The following section describes the process in which a real data set is examined for validity and accuracy.

3.2.1 Methods

For data collection, the data set of the study "Validity and Reliability of Consumer-Grade Optical Heart Rate Sensors to Assess Volume of Physical Activity and to Categorize Its Intensity" provided by Priv.-Doz. Dr. Dr. med. Mahdi Sareban, the Salzburger Landeskliniken and the Ludwig Boltzmann Institute for Digital Health and Prevention was used [98]. The data set collected ECG measurements and HR estimates of a wearable sensor (Garmin venu2s) via PPG. The data was collected over 24 hours in 32 subjects (66% male). 11 of these subjects were taking HR modifying drugs (β-blocker, ivabradine) during the time of the study. The data was collected with a sampling frequency of 1 Hertz (Hz), so the resulting HR value describes the number of Beats per Minute. The PPG recordings were extracted using the device's own software. Heart rate from the 4-lead Holter ECG device (Amedtec ECGpro, Aue, Germany) was exported from ECGpro, imported into GNU Octave (Copyright © 1998-2021 John W. Eaton) and transformed into 1 Hz format. Both extracted measurements were synchronized based on their time stamps, so that there was both a HR value from the ECG and a value from the wearable for each second.

3.2.2 Data Preprocessing

Despite the timestamp synchronization, other studies have also discussed that asynchronous results can occur, leading to errors and biased results. Reasons for these delays may include misaligned time stamps or systematic patterns. Delays can even be caused by the small time delays between the actual heartbeat (measured with the ECG) and the change in vessels in the extremities (measured with the wearable). Since the comparison of the measurement is done every second, a possible temporal shift can lead to a large bias in the results [15, 36, 64, 142]. Various studies describe the synchronization process in their investigation, using either an automated method or manual correction [44, 56, 63, 65, 66, 68, 81, 99, 103, 113, 115, 117, 132]. However, manual correction and visual inspection in this process is very time consuming and prone to potential errors. Therefore, Mühlen et al. [102] propose an automated method, such as shifting to the minimum Root Mean Squared Error (RMSE) or to the maximum Cross-Correlation-Coefficient [34, 129]. Coackley et al. [34] obtained the best results with this maximum Cross-Correlation-Coefficient shift.

To determine whether there was a systematic error in the PPG measurements in this data set, the measurements of each subject were examined and shifted to the maximum Cross-Correlation-Coefficient between the two HR measurements. The synchronized data sets were then analyzed. Table 3.1 shows a summary of the subjects' HR data.

	ECG	wearable
absolute number of HR measurements	2850230	2850230
HR mean	72.46 (23.67)	74.43 (18.40)
HR max	2143	201
HR min	0	33

Table 3.1: Summary of measurement of HR data of ECG and wearable
HR mean ... mean of all HR measurements ± standard deviation
HR max ... maximal HR value
HR min ... minimum HR value

3.2.3 Descriptive Analysis

The HR data was analyzed using a descriptive analysis. The intent of the descriptive analysis was to investigate the validity of the estimates of the wearables compared to the ECG recording. The data was processed and analyzed using Python.

Because the influence of exercise intensity was found to be highly relevant in the previous literature review, the HR data were additionally divided into four intensity ranges and analyzed within each range. Medical societies recommend relative exercise intensities based on physiological data such as percentage of Maximum Heart Rate (HRmax) to categorize four intensity ranges: **minor** (<57% HRmax), **light** (57-63% HRmax), **moderate** (64-76% HRmax) and **high** (>77% HRmax) [119].

Before the data sets were cleaned for further analyses, the HR data was analyzed for recording errors. A value was classified as an error if it was below the previously defined individual minimum HR (HRmin) or above the previously defined individual HRmax of the subject.

recording error	error$_{min}$	$\sum_i^N HR_i < HR_{min}$
	error$_{max}$	$\sum_i^N HR_i > HR_{max}$
	error$_{all}$	$\sum_i^N (HR_i < HR_{min}) + (HR_i > HR_{max})$

Across all intensity ranges, the ECG shows more errors (4031.69 ± 3768.65) compared to the wearable (252.37 ± 871.11) which can be seen in Table 3.2. Both the ECG and the wearable show more errors in this, recording heart rate as too low.

	ECG	wearable
error$_{all}$	4031.69 ± 3768.65	252.38 ± 871.11
error$_{min}$	3985.66 ± 3751.54	222.91 ± 875.41
error$_{max}$	46.03 ± 54.08	29.47 ± 63.27

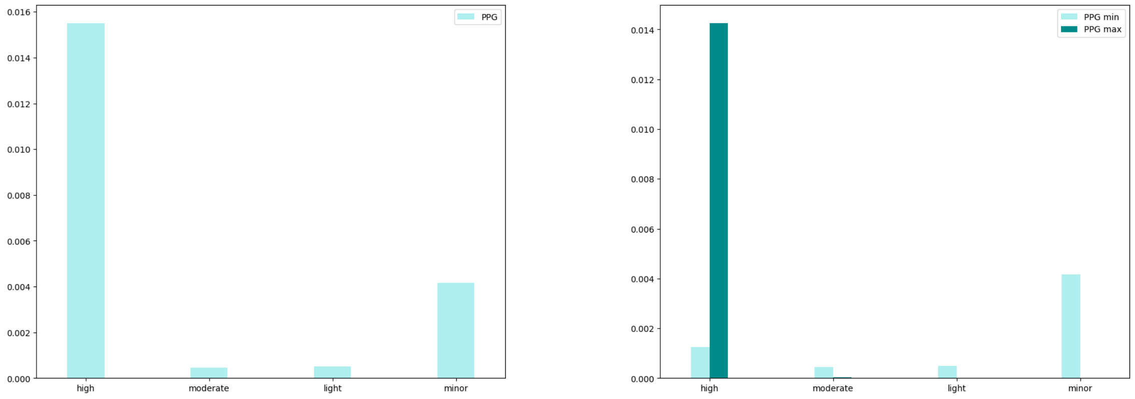
Table 3.2: Errors in ECG and wearable measurements: mean ± standard deviation

If the measurements are divided into the intensity ranges, the wearable tends to overestimate the HR in the high range and underestimate the HR in the minor range. There occur less errors in the light and moderate range, however also with a higher proportion of underestimating values. The most errors occur in the minor intensity range. Relative to the number of timestamps in the intensity range, the high range shows the most errors. Table 3.3 shows the absolute values of the errors while Figure 3.1a shows the relative values of $error_{all}$ compared to the absolute number of timestemps. The relative values of $error_{min}$ and $error_{max}$ are shown in Figure 3.1b.

Subsequently, all errors of the ECG were removed and the dataset could thus be used as a clean dataset for further analyses.

		wearable
range minor	absolute number	1728643
	error$_{all}$	206.16 ± 795.67
	error$_{min}$	206.13 ± 795.67
	error$_{max}$	0.03 ± 0.18
range light	absolute number	807366
	error$_{all}$	11.31 ± 52.28
	error$_{min}$	11.31 ± 52.28
	error$_{max}$	0
range moderate	absolute number	239206
	error$_{all}$	3.09 ± 15.54
	error$_{min}$	2.91 ± 15.54
	error$_{max}$	0.19 ± 1.06
range high	absolute number	75015
	error$_{all}$	31.81 ± 63.53
	error$_{min}$	2.56 ± 14.31
	error$_{max}$	29.25 ± 63.1

Table 3.3: Recording errors of wearable measurement in different intensity ranges



(a) $error_{all}$

(b) $error_{min}$ and $error_{max}$

Figure 3.1: Recording relative number of errors of wearable (PPG)

3.2.4 Results of Descriptive Analysis

Table 3.4 shows the mean, the standard deviation (std), the minimum (min) HR value, the maximum (max) HR value and the absolute number of heartbeats before and after the cleaning process. Over all intensity ranges, the mean of the ECG rises while the standard deviation decreases. The mean rises in all intensity ranges except in the minor range while the standard deviation decreases in all four intensity ranges.

		ECG		wearable	
		before	after	before	after
all ranges	absolute number	2850230	2631106	2850230	2631106
	mean	72.46	74.87	74.43	74.54
	std	23.67	18.48	18.4	17.72
	min	0	37	33	37
	max	2143	190	201	185
range high	absolute number	75015	66716	75015	66716
	mean	128.75	126.38	120.04	118.33
	std	23.92	20.16	24.66	23.78
	min	93	93	38	45
	max	2143	190	201	185
range moderate	absolute number	239206	227877	239206	227877
	mean	100.69	100.00	98.05	97.30
	std	14.91	14.86	15.40	15.21
	min	77	77	39	39
	max	144	144	169	169
range light	absolute number	807366	771150	807366	771150
	mean	83.58	83.21	82.80	82.40
	std	12.11	12.20	12.23	12.28
	min	62	62	36	41
	max	119	119	185	185
range minor	absolute number	1728643	1565363	1728643	1565363
	mean	60.91	64.90	65.15	65.48
	std	18.99	10.62	11.89	11.24
	min	0	37	33	37
	max	95	95	180	180

Table 3.4: Comparison of data before and after the cleaning process.

For the descriptive analysis, the following statistics were used to check the validity of the wearable's HR estimations (ppg) compared with the ground truth (ECG): Mean Absolute Error (MAE), Mean Directional Error (MDE), Root Mean Squared Error (RMSE) and Pearson Correlation Coefficient (corr). In addition, the relative number of matching beats was counted. A beat was classified as a match if the absolute difference between the HR values from the wearable and the ECG did not exceed three beats. The exact formulas are shown below, where HR_{ecg} refers to the measurement of the ECG, HR_{ppg} refers to the measurement of the wearable and N refers to the absolute number of timestamps.

$$\text{Mean Absolute Error:} \quad \text{MAE}(HR_{ecg}, HR_{ppg}) = \frac{\sum_{i=0}^{N-1} |HR_i^{ecg} - HR_i^{ppg}|}{N}$$

$$\text{Mean Directional Error:} \quad \text{MDE}(HR_{ecg}, HR_{ppg}) = \frac{\sum_{i=0}^{N-1} HR_i^{ecg} - HR_i^{ppg}}{N}$$

$$\text{Root Mean Squared Error:} \quad \text{RMSE}(HR_{ecg}, HR_{ppg}) = \sqrt{\frac{\sum_{i=0}^{N-1} (HR_i^{ecg} - HR_i^{ppg})^2}{N}}$$

$$\text{Pearson's Correlation Coefficient:} \quad \text{corr} = \frac{\sum_{i=1}^{N} (HR_i^{ecg} - \overline{HR^{ecg}})(HR_i^{ppg} - \overline{HR^{ppg}})}{\sqrt{\sum_{i=1}^{N} (HR_i^{ecg} - \overline{HR^{ecg}})^2} \sqrt{\sum_{i=1}^{N} (HR_i^{ppg} - \overline{HR^{ppg}})^2}}$$

$$\text{Matching Beat:} \quad \text{match} = \frac{\sum_{i=1}^{N} |HR_i^{ecg} - HR_i^{ppg}| \leq 3}{N}$$

Table 3.5 shows the results of the error metrics (MAE, MDE, RMSE), Pearson Correlation Coefficient (corr) and the Matching Beat (match) analysis, separated in the four intensity ranges. In the Mean Absolute Error, the Mean Directional Error, and the Root Mean Squared Error, the high intensity range provides the largest value and deviation. The minor intensity range shows the lowest error metrics and also the lowest deviation of errors. At the same time, the minor intensity range shows the largest value in the Pearson's Correlation Coefficient and the most Matching Beats. These are the lowest in the high intensity range.

	high	moderate	light	minor
MAE	9.022 ± 6.973	5.589 ± 3.796	3.900 ±2.343	2.129 ± 1.172
MDE	7.597 ± 7.229	3.677 ± 3.881	1.355 ± 1.758	-0.608 ± 0.580
RMSE	15.741 ± 9.116	9.501 ± 5.255	6.171 ± 3.212	3.942 ± 2.599
corr	0.547 ± 0.253	0.516 ± 0.190	0.666 ± 0.173	0.800 ± 0.194
match	0.648 ± 0.198	0.661 ± 0.184	0.697 ± 0.186	0.861 ± 0.118

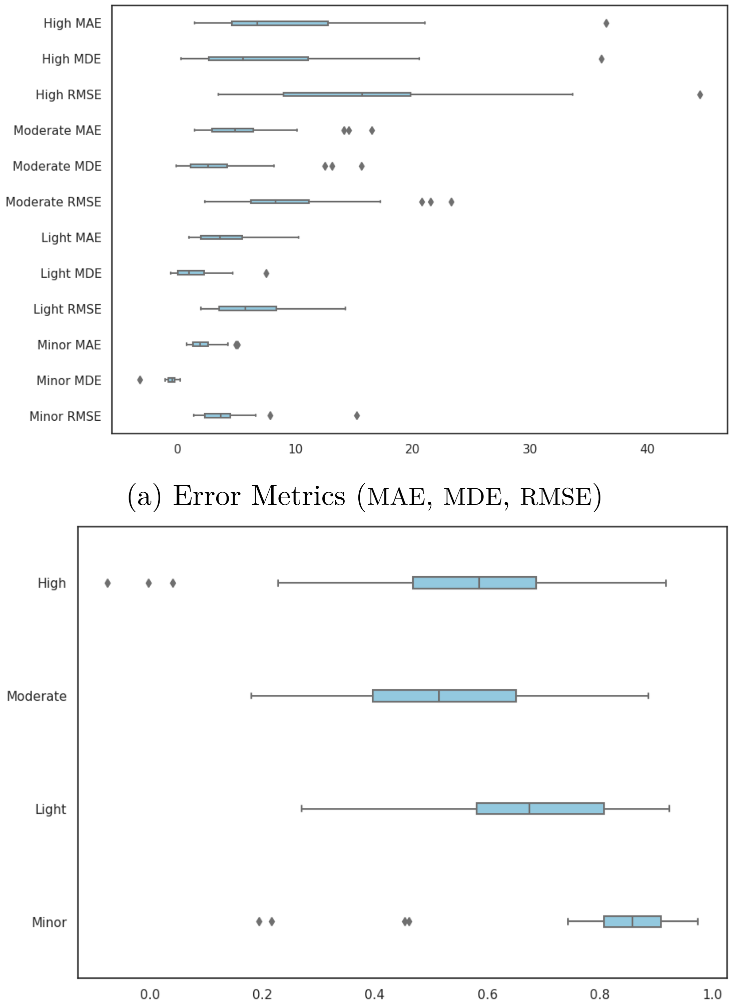
Table 3.5: Mean ± standard deviation of error metrics of all subjects

The boxplots of the error metrics separated in the intensity ranges are shown in Figure 3.2a, Figure 3.2b shows the Pearson's Correlation Coefficient.

In Table 3.6, the total value summarized over all subjects of MAE, MDE and RMSE can be seen. In all three error measurements, the high intensity range provides the largest values.

	high	moderate	light	minor
MAE	9.438	4.768	3.111	2.370
MDE	8.050	2.702	0.806	-0.577
RMSE	18.059	9.108	5.751	5.198

Table 3.6: Total Values of Error Metrics



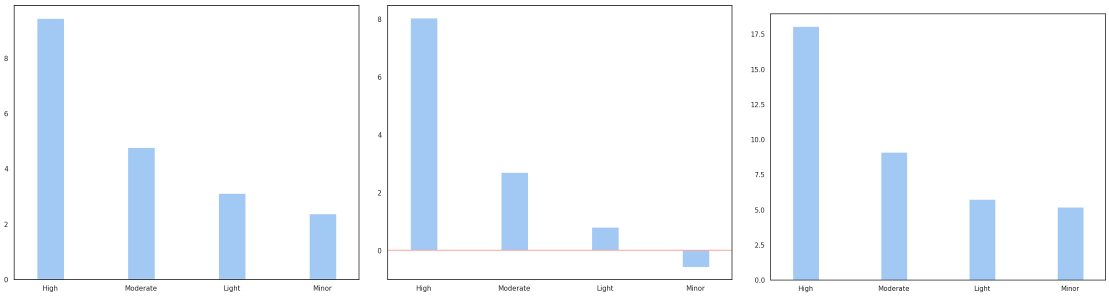
(a) Error Metrics (MAE, MDE, RMSE)

(b) Pearson's Correlation Coefficient

Figure 3.2: Boxplot of error metrics and Pearson's Correlation Coefficient between ECG and wearable separated into different intensity ranges

As shown in Figure 3.3b, the Mean Directional Error was negative for the minor intensity range, but positive for the light, moderate and high intensity ranges. This suggests that the wearble tends to overestimate the HR in the minor intensity range. In the other intensity ranges, the wearable tends to underestimate the HR, so the Mean Directional Error shows positive results. When the direction of the mean error is no longer considered, the Mean Absolute Error for the highest intensity range shows the largest value (Figure 3.3a). The Root Mean Squared Error is also largest in the high intensity range (Figure 3.3c).

The total value of the Pearson's Correlation Coefficient between the ECG and the wearable as well as the Matching Beats can be seen on Table 3.7. Consistent with the previous results, the high intensity range shows the lowest value. Here, the light



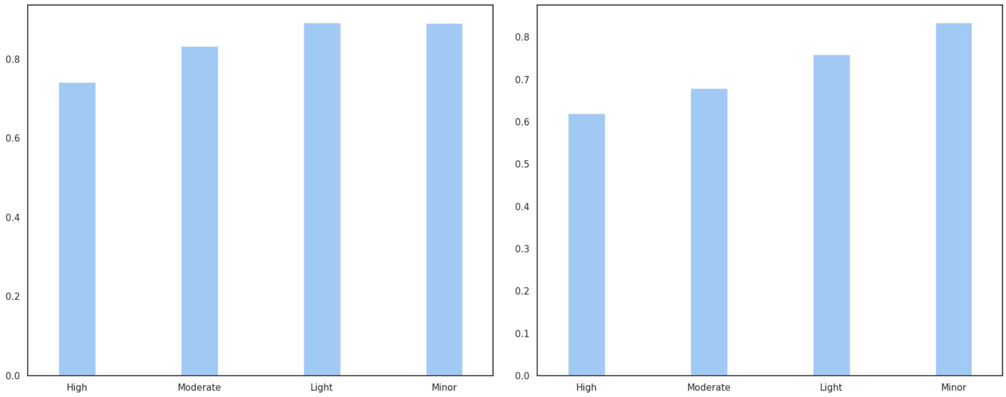
(a) Mean Absolute Error (b) Mean Directional Error (c) Root Mean Squared Error

Figure 3.3: Total values of error metrics of different intensity ranges

intensity range and the minor intensity range have the highest Pearson's Correlation Coefficient, therefore, the measurements of the wearable correlate best with the ECG measurement, which can also be seen in Figure 3.4a. Consistent with the previous results, the highest intensity range also shows the fewest Matching Beats (3.4b).

	high	moderate	light	minor
corr	0.741	0.833	0.892	0.890
match	0.619	0.679	0.760	0.834

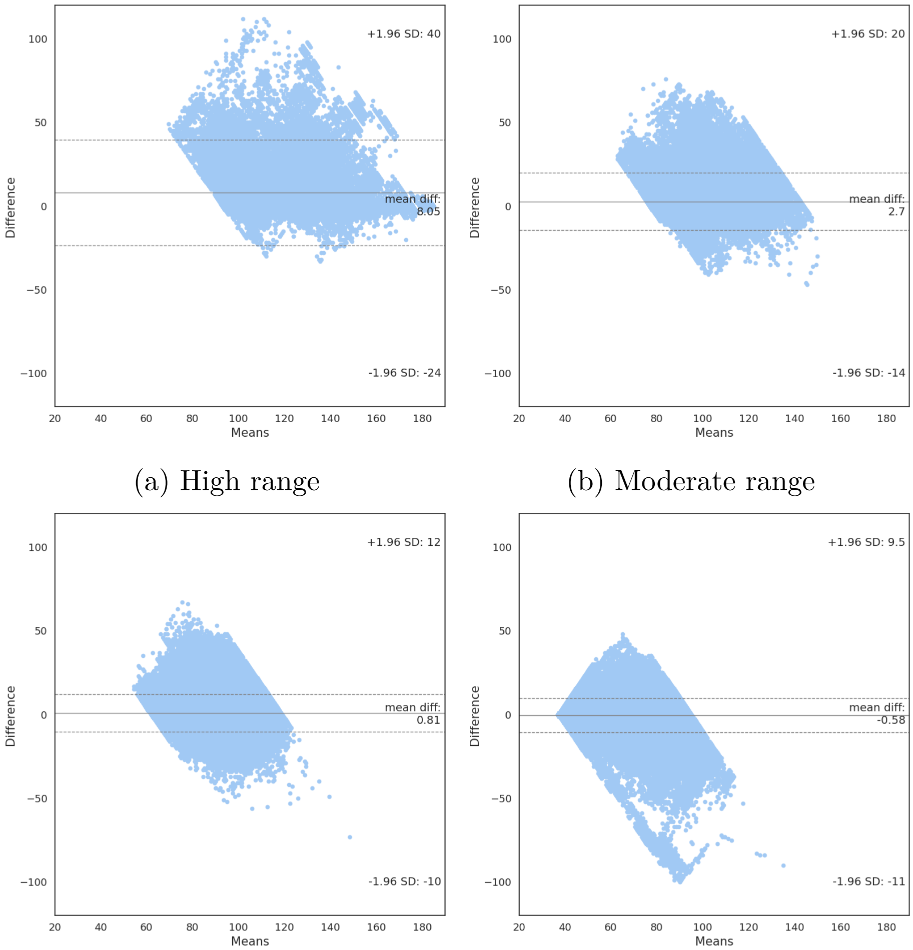
Table 3.7: Total values of Pearson's Correlation Coefficient (corr) and Matching Beats (match)



(a) Pearson's Correlation Coefficient (b) Matching Beats

Figure 3.4: Total values in different intensity ranges

The Bland-Altmann plots with respect to the four intensity ranges are shown in Figure 3.5. Again, it is visible that the highest intensity range has the largest deviation. The minor range has a larger difference towards the negative range which is consistent to the previous results.



(a) High range

(b) Moderate range

(c) Light range

(d) Minor range

Figure 3.5: Bland Altmann Plots of separate ranges

The Bland Altmann values are listed in Table 3.8.

	mean diff	sd diff	upper limit	lower limit
high	8.050	16.165	39.734	-23.634
moderate	2.702	8.698	19.751	-14.346
light	0.806	5.694	11.967	-10.355
minor	-0.577	5.166	9.549	-10.702

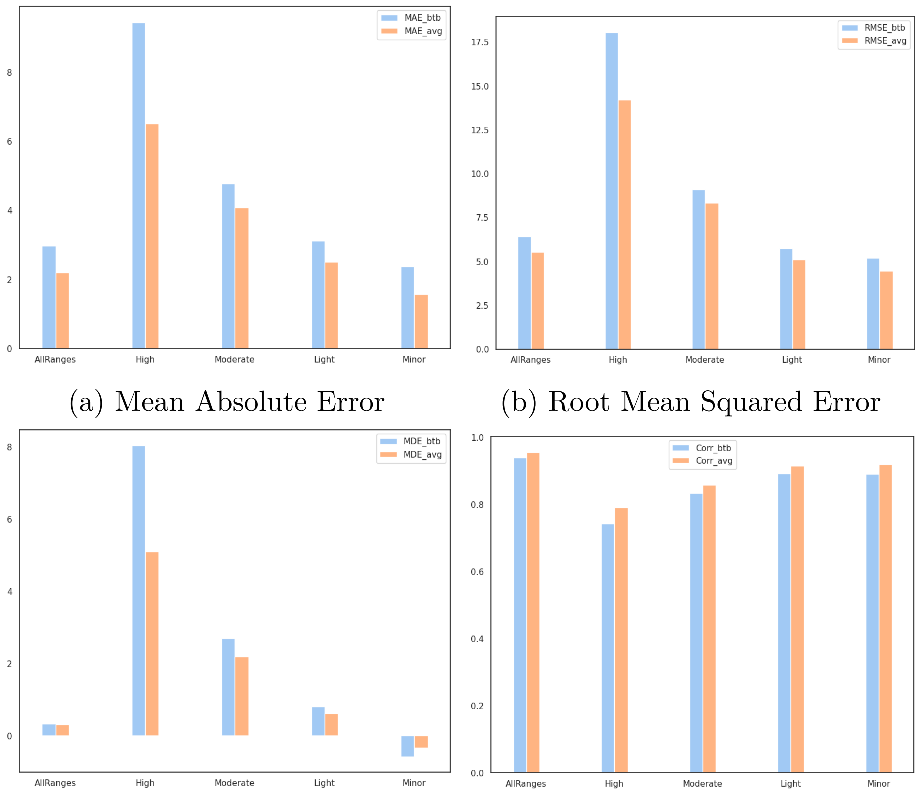
Table 3.8: Bland Altmann Plot values of different ranges

In addition to the Beat-to-Beat (btb) analysis, a 10 second Average Window (avg) analysis follows. For both the ECG and the wearable, the seconds were averaged over a rolling window of 10 seconds and then analyzed in terms of Mean Absolute Error, Root Mean Squared Error, Mean Directional Error and Pearson's Correlation Coefficient.

		btb	avg
all ranges	MAE	2.953 ± 1.397	2.215 ± 1.175
	MDE	0.307 ± 0.826	0.307 ± 0.825
	RMSE	5,826 ± 2,713	4.848 ± 2.825
	corr	0.927 ± 0.055	0.947 ± 0.052
high	MAE	9.022 ± 6.973	6.285 ± 6.847
	MDE	7.597 ± 7.229	4.866 ± 6.985
	RMSE	15.741 ± 9.116	10.699 ± 8.875
	corr	0.547 ± 0.253	0.729 ± 0.217
moderate	MAE	5.589 ± 3.796	4.934 ± 3.735
	MDE	3.677 ± 3.881	3.183 ± 3.822
	RMSE	9.501 ± 5.255	8.624 ± 5.357
	corr	0.516 ± 0.190	0.569 ± 0.201
light	MAE	3.900 ± 2.343	3.334 ± 2.177
	MDE	1.355 ± 1.758	1.080 ± 1.415
	RMSE	6.171 ± 3.212	5.610 ± 3.120
	corr	0.666 ± 0.173	0.702 ± 0.180
minor	MAE	2.129 ± 1.172	1.365 ± 0.949
	MDE	-0.608 ± 0.580	-0.297 ± 0.572
	RMSE	3.942 ± 2.599	2.885 ± 2.682
	corr	0.800 ± 0.194	0.893 ± 0.133

Table 3.9: Comparison of beat-to-beat (btb) vs 10 second average (avg)

In all three error metrics, averaging over 10 seconds shows a reduction in error. The difference between the beat-to-beat results and the averaged value over 10 seconds shows the same ranking, but the error values become smaller overall. This is evident for the MAE (3.6a), the RMSE (3.6b), and the MDE (3.6c). The Pearson's Correlation Coefficient, on the other hand, becomes larger for the averaged values (3.6d). Table 3.9 shows the results of this analysis.



(a) Mean Absolute Error

(b) Root Mean Squared Error

(c) Mean Directional Error

(d) Pearson's Correlation Coefficient

Figure 3.6: Comparison of btb versus avg of different error metrics over separate ranges

3.3 Results

To verify the validity of the wearable sensor, several descriptive statistics were used that analyzed the wearable in comparison to the gold standard the ECG. The wearable shows the most recording errors in the minor intensity range, however looking at the relative number of recording errors, the high intensity range shows the highest value. In the highest intensity range the wearable shows the largest errors (MAE, MDE and RMSE) as well as the smallest Pearson's Correlation Coefficient and the less Matching Beats. Comparing the measurements not beat-to-beat, but over an averaging window of 10 seconds, all error metrics can be reduced and the Pearson's Correlation Coefficient increases. Nevertheless, the individual intensity ranges show up in the same order measured by the magnitude of the error.

These results are consistent with previously obtained findings from the literature, where the wearables also showed the greatest inaccuracy in the high intensity ranges [24, 40, 43, 46, 68, 105, 125, 153, 176].

4 Prediction of Measurement Error

AI and Machine Learning have been used to develop models that can predict disease risk, treatment effectiveness, and the outcome of medical procedures. For example, Machine Learning models have been used to predict the risk of heart attack, the effectiveness of cancer treatments and the outcome of surgery. These models have helped physicians make better decisions about patient care. Natural language processing has been used to extract information from medical texts. This information can be used to improve the diagnosis and treatment of disease. Natural language processing has been used to extract information from medical records and to develop systems that can answer questions about diseases. These systems have helped physicians provide better care to their patients [11]. The aim now, is to develop a model which is able to capture and learn the measurement error that was confirmed in the investigating chapter and thus be able to predict occurring measurement errors.

Since physiological data, such as the recording of HR, is also understood as a time series, it is necessary to apply the best methods proposed by the literature for time series prediction. A literature research was carried out at the beginning in order to be able to select a suitable method.

4.1 Literature Research: Time Series Prediction

In recent decades, interest regarding time series datamining has grown rapidly [4, 32, 87, 86, 92]. The components of time series datamining include pattern recognition, clustering, classification and prediction [150].

Prediction is the most common and important application of time series [165]. Time series prediction is called prediction by examining pattern from past data [101]. In recent years, a lot of research has been done to understand the future using time series data. [112]

Time series forecasting can be found in business and academia, in the financial sector, for forecasting index prices, stock closing prices, production revenue, sales volume

[60, 76, 85, 179], for forecasting electricity load [37, 96], for weather forecasting [110, 122] or in the field of medicine and health. Time series forecasting is used in the medical field in many different situations, such as dialysis of critically ill patients, predicting mortality risk in pediatric intensive care, predicting hypotension episodes in critical care or predicting morbidity of tuberculosis [3, 47, 84, 111]. Arima, Linear Regression, Support Vector Regression (SVR), Gradient Boosted Regression Tree (GBRT) and Multilayer Perceptron (MLP) are classical regression models [128, 141] which can perform efficient numerical prediction and time series forecasting. These regression models are easy to implement and also require less computation time. However, these models have two limitations: they must use fixed-length features and cannot fully exploit sequential dependence [89]. This has sparked interest in developing more powerful forecasting methods and interest is quickly shifting toward neural forecasting methods [168].

Deep learning models have become a promising tool for time series prediction [16, 112]. By "learning without assumptions," neural networks have distinct advantages in real-world applications where data is easy to collect while relationships are difficult to discern. Neural networks are non-parametric, non-linear and have universal functional approximations which can learn and capture relationship from the data [16]. They have the advantage and strength of being able to automatically learn temporal dependencies and automatically process temporal structures such as trends and seasonality. They can automatically learn arbitrary complex mappings of inputs to outputs and support multiple inputs and outputs [21]. The success of the Machine Learning processes depends on how well the features were extracted, which is a subjective process and often leads to overfitting. Furthermore, Deep Learning methods can speed up this feature extraction process, which can be complicated and time consuming. Thus Deep Learning methods provide better results and offer better generalization. [108]

The most used Deep Learning algorithms are Convolutional Neural Networks (CNNs), Long Short Term Memory Networks (LSTMs), Recurrent Neural Networks (RNNs), Generative Adversarial Networks (GANs), Deep Belief Networks (DBNs), Restricted Boltzmann Machines (RBMs) or Autoencoders.

There are several studies in which the prediction of time series using CNN was successful. Livieris et al. [91] used a CNN-based model to predict the price of gold. Gamboa et al. [55] compared several Deep Learning models for time series prediction, including CNNs, and found that CNNs produced better results. Fawaz et al. [49] compared several CNN-based models for time series classification and found that they outperformed other Machine Learning algorithms. Zhao et al. [177] investigated the classification accuracy of a novel CNN model, which outperformed conventional methods. In addition, they showed that Deep Learning methods detect and learn more robust features and therefore perform better on classification problems. Warrick et al [163] developed a method for short-term HR prediction of perinatal fetal HR using LSTM. In addition, they proposed a model combining CNN and LSTM to avoid the weakness of LSTM, which performed better. Zhang et al. [176] also developed a Deep Learning network to classify multivariate time series and achieved significantly better results. For this purpose, they modified a CNN model and combined it with an MLP. Applications of CNNs in the health domain include ECG classification [79], structural health monitoring [1] and motor disturbance detection [67]. These techniques have been shown to be very effective in capturing long-term dependencies and nonlinear dynamics [97].

Since their introduction in the 1990s, CNNs have contributed enormously to the success of Deep Learning. CNNs are designed to mimic the way the human brain thinks. They learn fully automatically, which allows them to extract features that are salient in the input data across different layers. [123]

There are several studies that investigated different algorithms and models using Machine Learning and Deep Learning to predict time series of HR. All studies showed better results for Deep Learning algorithms compared to traditional Machine Learning algorithms. Since CNNs can completely solve nonlinear problems for a large amount of data to some degree, it is often used for the prediction of physiological parameters [140, 144, 149]. However, CNNs have a weakness of not being able to learn the variation of peak features when the data have high volatility and instability [175].

Masum et al. [97] investigated different Deep Learning models such as LSTM, BI-LSTM, and CNN for predicting Blood Pressure (BP) and Heart Rate (HR) from univariate and multivariate time series. Here, they developed models for predicting blood pressure and HR 30 minutes ahead (univariate) and BP+HR multivariate, respectively (resp). The multivariate version showed better results. Murugesan et al [104] used a CNN-based model to classify ECG signals into different types of arrhythmias that may affect HR. Niu et al [108] developed a CNN-based model to classify ECG signals into different HR categories. Seong-Hyun et al. [77] used a CNN-based model to estimate HR from PPG signals. Qiu et al. [121] used CNN to calculate HR from facial videos. Reiss et al. [126] estimated HR from PPG and accelerometer data using a Fourier Transform (FFT) and a four-layer CNN model with preprocessing by z-normalization. Biswas et al [18] proposed a four-layer deep neural network with two CNN layers and two LSTM layers to improve the accuracy of HR disease in naturalistic measurements. The proposed system improved Mean Absolute Error accuracy in HR prediction on their dataset of 22 PPG recordings. Zhang et al. [175] set the goal to predict HR according to the current real-time HR measurement to effectively predict and prevent cardiovascular diseases. For this purpose, they developed a model combining CNN and Gated Recurrent Unit (GRU) to exploit their respective advantages. Experimental results showed that this model exhibited higher prediction accuracy than other traditional methods. Brophy et al. [20] applied two Deep Learning methods, one for human activity detection, one for

HR estimation. For HR estimation, they used a publicly available PPG dataset that captured HR at a rate of 256 Hz for 10 minutes. The PPG recording was compared with a simultaneously running ECG, which served as the ground truth. Before recording the PPG signal, it was down sampled to different sampling frequencies and a classifier was trained using the recalculated frequencies (30 Hz, 15 Hz, 5 Hz, 1 Hz). In the process, they developed a CNN with a regression layer as the output layer, named CNNR (Convolutional Neural Network with Regression). This is a four-layer one-dimensional network with batch normalization and ReLUs. To compare the performance of the CNNR model, an open source toolkit was used, which is also used for HR estimation of PPG data. To achieve the best result, the parameters of the CNNR model were chosen. They obtained comparable results to the toolkit and additionally discovered that the CNN model obtained better results at low recording frequency (10 Hz) than at 15 or 30 Hz. The reasons for this have not yet been fully investigated. Taye et al [152] used different Machine Learning algorithms to predict the occurrence of imminent ventricular tachyarrhythmia. They compared CNN, ANN, Support Vector Machine (SVM), K-Nearest Neighor (KNN), with CNN outperforming the other algorithms. Shyam et al [139] were the first to attempt to estimate PPG signals independent of patient and data set. They compared different Deep Learning Methods such as CNN, LSTM and Fully Conventional Network (FCN) on two different datasets and achieved satisfactory results with their approach (MAE 3.36 ± 4.1 average error). It confirms that the presented CNN model can adapt to PPG devices and needs little training to adapt to a new device with a new hardware. They worked with a window size of 8 seconds with 6 seconds overlap and a sampling frequency of 125 Hz.

In recent years, in the field of Deep Learning, the transformer model has also attracted much attention due to its excellent performance in various domains, such as natural language processing, computer vision and language processing [38, 42, 165]. Due to the advantage that transformers have with great modeling abilities, they can capture dependencies and interactions in sequential data well. They are also interesting in the field of time series modeling, such as in prediction, anomaly detections, and also for classification problems [85, 157, 165, 169, 171, 172, 173]. Li et al [85] compared transformers in univariate time series forecasting and showed that this outperformed conventional Machine Learning. Zerveas et al. [173] presented a transformer-based framework for unsupervised learning of multivariate time series representation. Using evaluations on several benchmark datasets, they showed that this modeling approach outperforms all existing supervised state-of-the-art methods. Liu et al [90] combined transformers with gating mechanisms. This combination shows competitive performance on time series classifications in various experiments compared to state-of-the art Deep Learning models. Yang et al. [172] implemented an approach using transformers to reprogram a pre-trained acoustic model for time series classification. They achieved at least as good of results as state-of-the-art techniques. Cai et al. [26] used transformers to predict traffic and capture spatiotemporal dependencies. They achieved the best results with a time series segment method in combination with a neural graph convolutional network and transformer.

In the field of medicine and health, transformers have been studied very little. Che et al. [31] attempted to use transformers to classify an ECG signal based on arrhythmia. They presented a model that combines CNN and a transformer for ECG signal classification. By combining them, they achieved significantly improved performance. Katrompas et al. [73]compared an LSTM with Self Attention model and a transformer model for classification and prediction of time series, where the LSTM with Self Attention gives better results.

Thus, a combination of CNN and Transformers show great potential for predicting HR data. This combination is going to be compared to a simple CNN model. Additional a conventional Linear Regression model will be compared to those two Deep Learning approaches.

4.2 Methods

The goal of this work is to train a model to recognize the current measurement error of the wearable and predict the difference between the wearable's estimate and the ground truth's correct HR measurement representing the actual heartbeat. For this purpose, three models are compared: Multiple Linear Regression, a Deep Learning Model using CNN and a Deep Learning Model using a combination of CNN and Transformers. For all approaches, the data was first preprocessed and put into a format so that it could be subsequently be used for prediction. After that, the data was divided into training and test set to later evaluate the performance of each model and compare them.

4.2.1 Models

There are three models, which will be used for this work. The first one is a Linear Regression Model, the second and the third are Deep Learning models.

Linear Regression Model

As mentioned in the previous section, there are simple methods to predict the future from past data. One basic method is Linear Regression. Linear Regression is a statistical technique that studies and models the relationship between variables. It is used in many fields such as engineering, physical and chemical sciences, and life and biological sciences. It may be the most commonly used statistical technique. Regression is used for data description, parameter estimation, prediction and estimation and control, among others [127]. This Linear Regression model is intended to serve as the basic model in this thesis, since it involves less computational effort and thus will serve as a comparative model.

Deep Learning

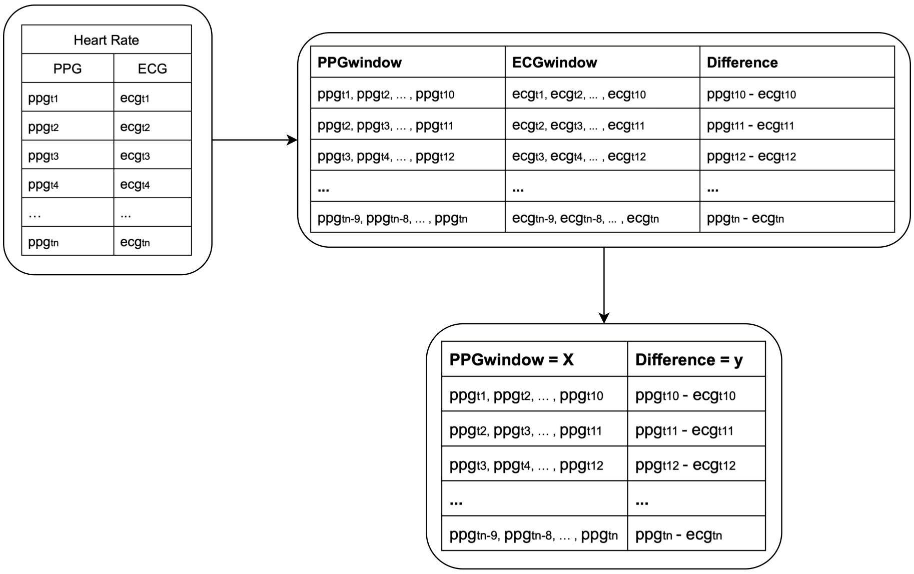
In the literature review, both CNN and Transformers show great promise. CNNs have been used to work with heart data however Transformers have not been studied much in the field of medicine and health. There are already studies combining CNN and Transformers to classify ECG, but not yet to predict PPG signals. Based on this, a simple CNN model and a CNN in combination with Transformer, referred to as CnnTrans in the following sections, are used in this work

4.2.2 Data Preprocessing

The data was prepared in such a way that a prediction model could be applied subsequently. First, the data was cleaned of missing measurements. The ECG, which is considered as the ground truth, was additionally cleaned of obvious errors and outliers ($error_{all}$). Then the data was transformed into a supervised learning mode. In this mode, the idea is that the model recognizes and learns an association between the input and output variables. The model is later trained based on the input data (X) to predict the output data (y). To transfer the time series data into supervised learning, the input data was configured into an Observation Window (X). Thus, the PPG signal was divided into windows of a size of 10 seconds, with an overlap of 9 seconds with the previous window and a slide of 1 second. The output variables (y) described the difference between the HR estimate of the wearable to the HR measurement of the ECG for that specific window as a real value. The preprocessing into the window (X) and difference (y) is shown in the figure 4.1.

4.2.3 Data Splitting

The data was divided into 80% training set and 20% test set. The data of all 32 subjects contain a total of 2453269 windows. In order not to split the individual data sets of the subjects, subject 1 to 25 were used as training data and 26 to 32 as test data. This results in a split of 2120672 training values and 598596 test values. The models were trained and optimized using the training data and evaluated using the test data.



Heart Rate	
PPG	ECG
ppg_{t1}	ecg_{t1}
ppg_{t2}	ecg_{t2}
ppg_{t3}	ecg_{t3}
ppg_{t4}	ecg_{t4}
...	...
ppg_{tn}	ecg_{tn}

PPGwindow	**ECGwindow**	**Difference**
ppg_{t1}, ppg_{t2}, ... , ppg_{t10}	ecg_{t1}, ecg_{t2}, ... , ecg_{t10}	ppg_{t10} - ecg_{t10}
ppg_{t2}, ppg_{t3}, ... , ppg_{t11}	ecg_{t2}, ecg_{t3}, ... , ecg_{t11}	ppg_{t11} - ecg_{t11}
ppg_{t3}, ppg_{t4}, ... , ppg_{t12}	ecg_{t3}, ecg_{t4}, ... , ecg_{t12}	ppg_{t12} - ecg_{t12}
...	...	...
ppg_{tn-9}, ppg_{tn-8}, ... , ppg_{tn}	ecg_{tn-9}, ecg_{tn-8}, ... , ecg_{tn}	ppg_{tn} - ecg_{tn}

PPGwindow = X	**Difference = y**
ppg_{t1}, ppg_{t2}, ... , ppg_{t10}	ppg_{t10} - ecg_{t10}
ppg_{t2}, ppg_{t3}, ... , ppg_{t11}	ppg_{t11} - ecg_{t11}
ppg_{t3}, ppg_{t4}, ... , ppg_{t12}	ppg_{t12} - ecg_{t12}
...	...
ppg_{tn-9}, ppg_{tn-8}, ... , ppg_{tn}	ppg_{tn} - ecg_{tn}

Figure 4.1: Data preprocessing for supervised learning

Table 4.1 shows the number of windows of the training and test set. Additionally, it shows how many of these windows are located in the high, moderate, light or minor intensity range.

	whole dataset	high	moderate	light	minor
training data	2120654	52927	186868	623994	1256865
test data	598578	16468	50323	185371	346416

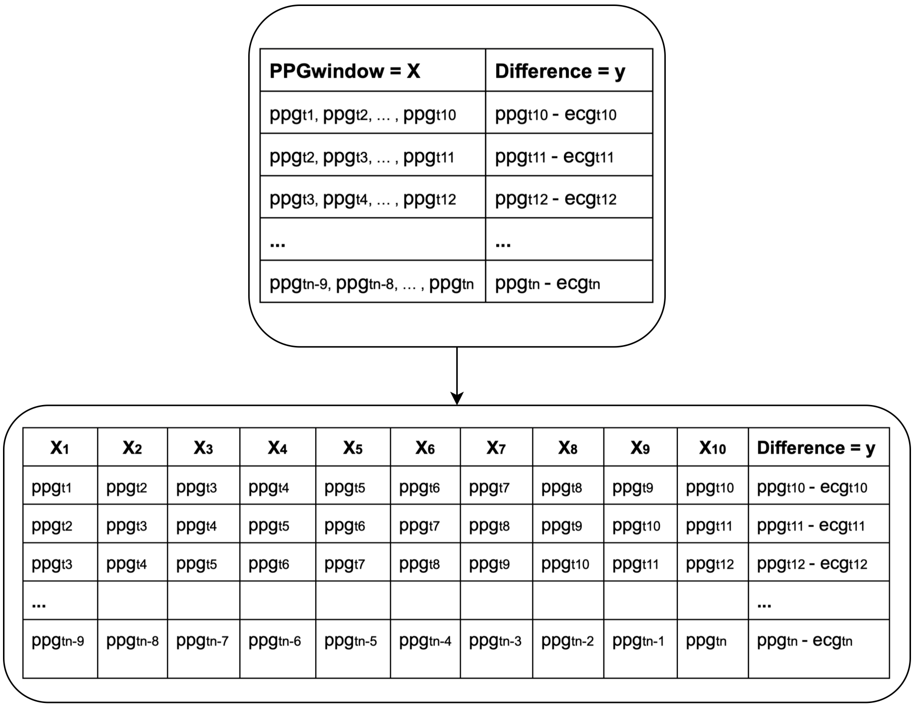
Table 4.1: Number of windows in training and test set

4.3 Model Training

In the following, the structure of each model will be explained in more detail as well as the progress of the training.

4.3.1 Multiple Linear Regression

For the linear models, the data was further processed by dividing the Observation Window into a feature matrix. Each column of the matrix represents a different value of the window. Each row of the matrix corresponds to one observation window. The model uses each time stamp of the window as a regressor variable. Figure 4.2 shows the preprocessing of the data for the regression model.



PPGwindow = X	Difference = y
ppg_{t1}, ppg_{t2}, ... , ppg_{t10}	ppg_{t10} - ecg_{t10}
ppg_{t2}, ppg_{t3}, ... , ppg_{t11}	ppg_{t11} - ecg_{t11}
ppg_{t3}, ppg_{t4}, ... , ppg_{t12}	ppg_{t12} - ecg_{t12}
...	...
ppg_{tn-9}, ppg_{tn-8}, ... , ppg_{tn}	ppg_{tn} - ecg_{tn}

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Difference = y
ppg_{t1}	ppg_{t2}	ppg_{t3}	ppg_{t4}	ppg_{t5}	ppg_{t6}	ppg_{t7}	ppg_{t8}	ppg_{t9}	ppg_{t10}	ppg_{t10} - ecg_{t10}
ppg_{t2}	ppg_{t3}	ppg_{t4}	ppg_{t5}	ppg_{t6}	ppg_{t7}	ppg_{t8}	ppg_{t9}	ppg_{t10}	ppg_{t11}	ppg_{t11} - ecg_{t11}
ppg_{t3}	ppg_{t4}	ppg_{t5}	ppg_{t6}	ppg_{t7}	ppg_{t8}	ppg_{t9}	ppg_{t10}	ppg_{t11}	ppg_{t12}	ppg_{t12} - ecg_{t12}
...										...
ppg_{tn-9}	ppg_{tn-8}	ppg_{tn-7}	ppg_{tn-6}	ppg_{tn-5}	ppg_{tn-4}	ppg_{tn-3}	ppg_{tn-2}	ppg_{tn-1}	ppg_{tn}	ppg_{tn} - ecg_{tn}

Figure 4.2: Preprocessing for Linear Regression model

The linear model was created using the scikit-learn library. The model takes the feature matrix (X_1 to X_{10}) and the target variable (y) as input and analyzes the relationship between those. It also assumes a linear relationship between the features and the target variable and assumes that the target variable can be expressed as a linear combination of the feature variables, with the coefficients assigned to each feature representing weights. By fitting the model, the coefficients indicating the strength and direction of the relationship between the characteristics are estimated.

4.3.2 CNN

A Deep Learning model using the Keras Deep Learning framework was built. The model consits of one input layer, four pairs of convolutional layers followed by batch normalization and ReLU activation. After that there is a global average pooling layer to reduce the spatial dimensions of the features. A flattening layer follows and two fully connected dense layers are added. The first dense layer has ReLU activation with 32 units, while the second layer has 1 unit. The model uses the Adam optimizer and a custom loss function, which is defined by the absolute difference between the true difference of the ECG and the wearable (=$diff_{true}$) and the predicted difference of the model (=$diff_{pred}$).

$$\text{loss function} = |diff_{true} - diff_{pred}| \tag{4.1}$$

After preliminary tests, ranges of possible values for hyperparameters are defined, such as epoch number = [50, 100, 200, 300], hidden neurons [8, 16, 32, 64], or kernel size = [3, 5, 7]. The optimal hyperparameters can be determined based on the prediction performance which were evaluated and identified using GridSearch and cross-validation. The optimal parameters are as follows: 200 epochs, 8 hidden neurons in the first two and 16 hidden neurons in the third and fourth layer with a kernel size of 3. The structure of the architecture of the whole Deep Learning process is shown in Figure 4.3. The CNN model architecture is shown in Figure 4.4a.

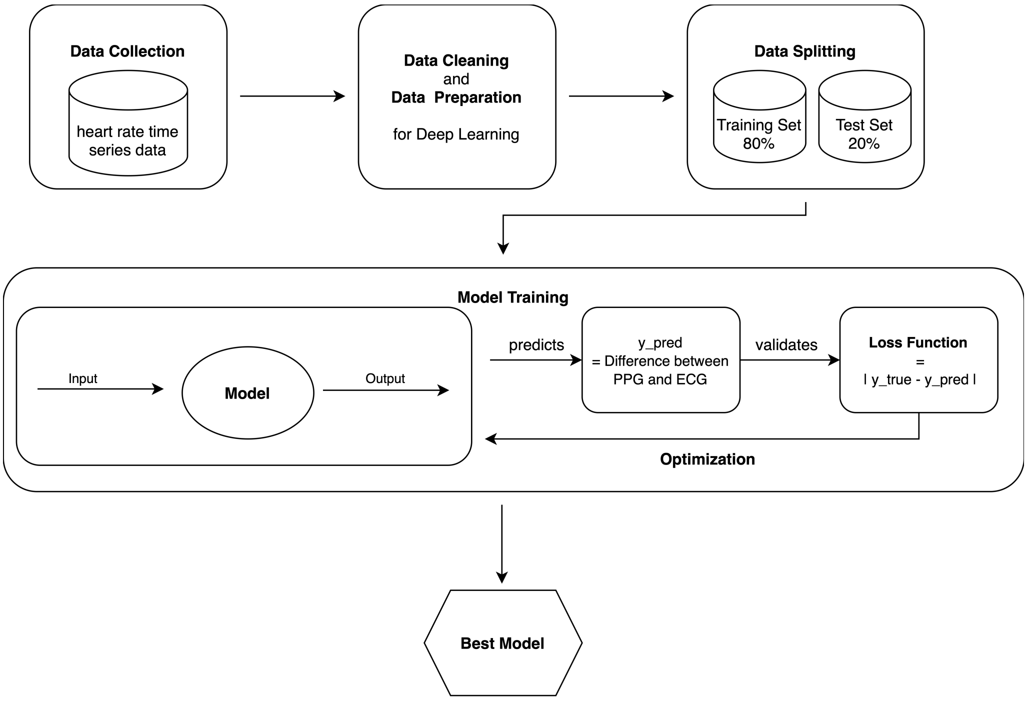
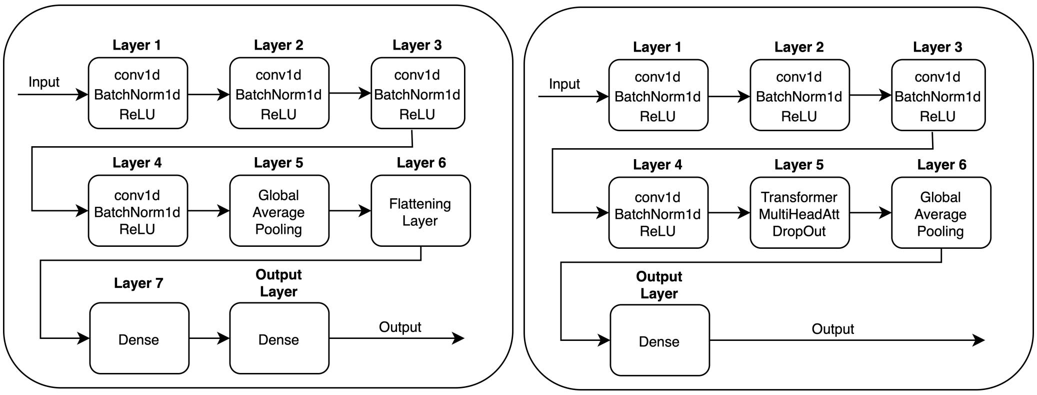

Data Collection
heart rate time series data
Data Cleaning and Data Preparation for Deep Learning
Data Splitting
Training Set 80%
Test Set 20%
Model Training
Input
Model
Output
predicts
y_pred = Difference between PPG and ECG
validates
Loss Function = I y_true - y_pred I
Optimization
Best Model

Figure 4.3: Architecture of Deep Learning model

4.3.3 CNN and Transformer

Using the Keras Deep Learning framework with Tensor Flow backend, a Deep Learning model was built. The model initially consists of four convolutional layers. Each layer creates a one dimensional convolutional layer and applies the ReLU activation function to its output. In addition, each layer applies batch normalization to the output. After these two convolutional layers, the model applies a multi-head self-attention mechanism. It splits the input into multiple heads, performs attention computations, and returns the attention output. Then, it undergoes a DropOut realization is performed to avoid overfitting. Batch normalization is applied on top of this, which helps stabilize the training process. Global average pooling is performed on the output of the transformer block. This calculates the average value for each channel along the time dimension, resulting in a fixed-length representation of the input sequence. Next, a fully connected dense layer is added and the ReLU activation function is applied. Finally, a dense layer is added, representing the output of the model. Overall, this defines a model architecture that consists of four convolutional layers, a transformer layer and fully connected layers. The model takes an input sensor, processes it through the defined layers then produces an output sensor. The same loss function, as used for the CNN (eq:4.1), was used. The optimal hyperparameters were determined using GridSearch and cross-validation and were as follows: 200 epochs, 8 hidden neurons in the first two and 16 hidden neurons in the third and fourth layers with a kernel size of 3. The model including the architecture is shown in Figure 4.4b.



(a) Architecture of CNN model

(b) Architecture of CnnTrans (CnnTrans) model

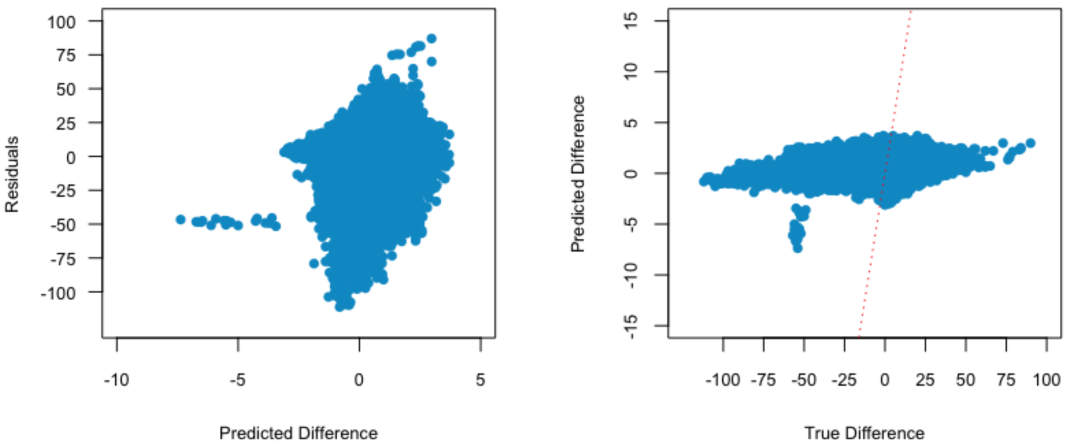
4.4 Results

The following section describes the results of applying the models to the test data set.

Using the training set as input data, the **Multiple Linear Regression** (MLR) model yields the following equation:

$$y = -2.33 - 0.10 * X_1 + 0.01 * X_2 + 0.02 * X_3 + 0.01 * X_4 + 0.01 * X_5 - 0.02 * X_6 \\ - 0.02 * X_7 - 0.01 * X_8 + 0.02 * X_9 + 0.11 * X_{10}$$

When evaluating with the test data, the model achieves a RMSE of 7.5753, a MAE of 3.7731 and a R^2 of 0.009257. The residuals as well as the predicted values versus the true values are plotted in Figure 4.5. It is clear that the model makes the most predictions close to 0, as can be seen in Figure 4.5b. It has the biggest problems with correctly predicting the values that fall further towards the extremes. The red line shows how the points should lie in the optimal case. This finding can also be seen in Figure 4.5a. Since most of the predictions are close to 0, the model therefore makes the largest errors in this area and shows the largest residuals.



(a) Predicted Difference vs Residuals (b) True Difference vs Predicted Difference

Figure 4.5: Multiple Linear Regression (MLR)

To understand the individual relationships between each independent variable and the dependent variable, plots were created. The plots of all independent variables appear to be very similar. Figure 4.6 shows X_1 versus the true difference as an example. The other plots of the independent variables are attached in the appendix.

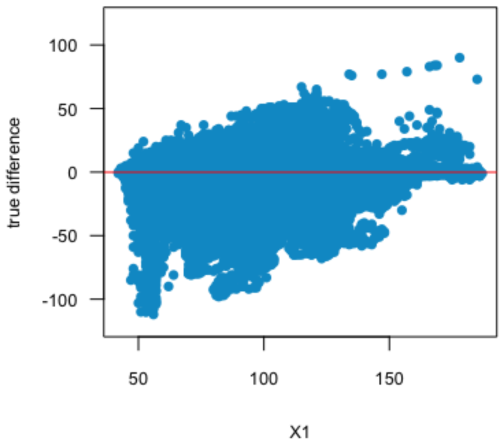


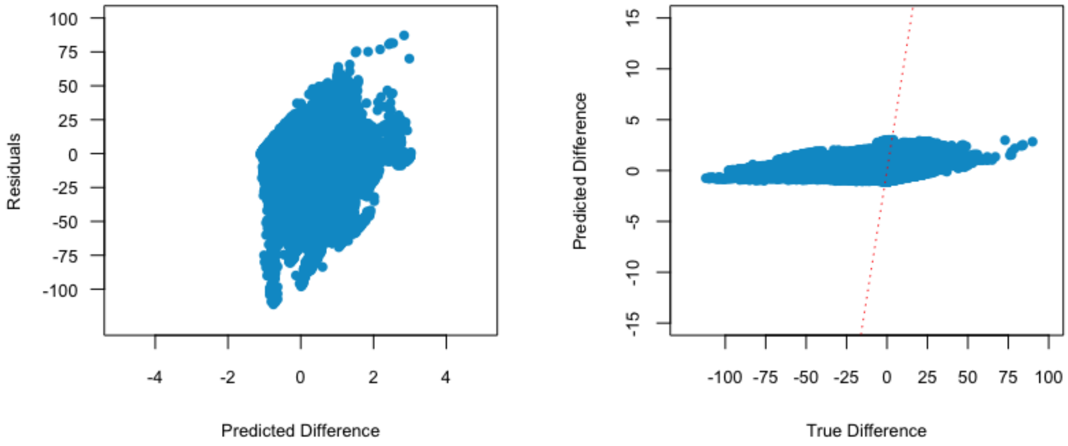
Figure 4.6: Independent variable X_1 to true difference in test data set

The plots show a weak linear trend, with the slope being slightly positive. This suggests that an increase in X_1 leads to an increase in the true difference. However, the range of X_1 values shows a tendency to underestimate by more than 100 in the lower range, and overestimate by about 25 to 50 in the upper range, even with a few outliers from 75 to 100. There is also different scatter with the deviation. The higher the values are, the lower the deviation gets.

Since the plots of the independent variables appear to be similar, Variance Inflation Factors (VIFs) were calculated to assess multicollinearity among the independent variables. The results showed that all of the VIFs were significantly high, suggesting that multicollinearity was a problem. This indicates that these variables are highly correlated with each other, which can make it difficult to estimate their individual effects on the dependent variable. To address the problem of multicollinearity, the average of each adjacent pair of variables was calculated. This resulted in five new variables, which were then used as the explanatory variables in a new **Multiple Linear Regression** model (MLR2). The following equation was used to represent the new model:

$$y = -2.33 - 0.11 * X_1 + 0.06 * X_2 - 0.02 * X_3 + -0.06 * X_4 + 0.16 * X_5$$

The data were thus re-trained with the new model and the difference predicted for the test data. The model achieved a RMSE of 7.5737, a MAE of 3.7586, and a R^2 of 0.009206. Figure 4.7b shows that the model still has a tendency to predict the values close to 0. The largest residuals appear also in those cases and seem to get smaller when the predicted difference is larger, with some outlier. This can be seen in Figure 4.7a.



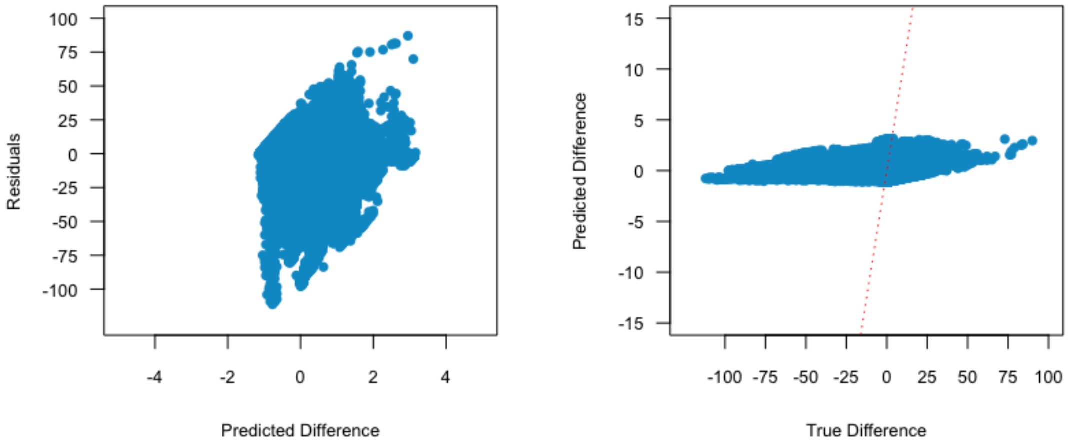
(a) Predicted Difference vs Residuals (b) True Difference vs Predicted Difference

Figure 4.7: Adapted Multiple Linear Regression (MLR2)

VIFs were also calculated for this model, which were also significantly high. To address the issue of multicollinearity, a third variant of the original MLR model was created. In this model, the variable with the highest VIF was removed iteratively until all VIFs were below the threshold of 10. This procedure resulted in a model with only one independent variable, namely:

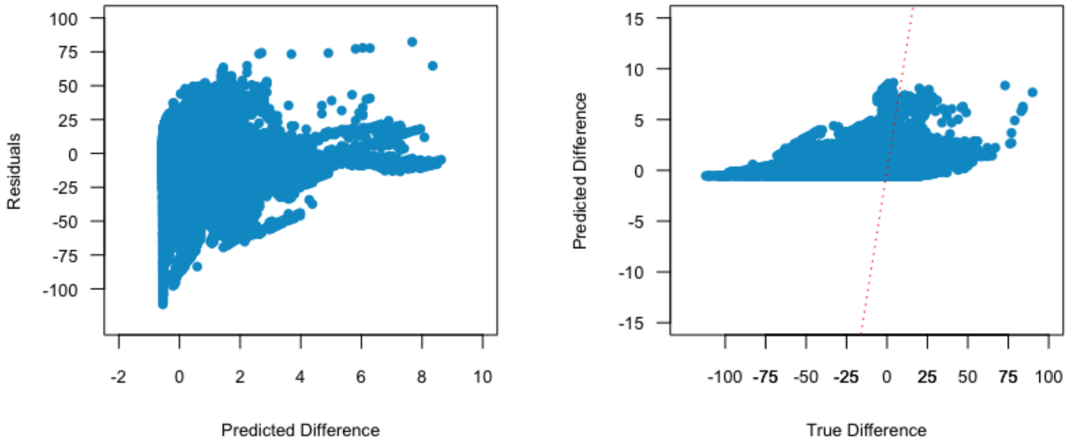
$$y = -2.41 + 0.03 * X_{10}$$

The new **Univariate Linear Regression** (ULR) achieves a RMSE of 7.5750, a MAE of 3.7613, and an R^2 value of 0.007778. Figure 4.8a plots the residuals against the predicted differences. Figure 4.8b shows the true difference versus the predicted difference. In both plots the remaining problem can be seen.



(a) Predicted Difference vs Residuals (b) True Difference vs Predicted Difference

Figure 4.8: Univariate Linear Regression (ULR)



(a) Predicted Difference vs Residuals (b) True Difference vs Predicted Difference

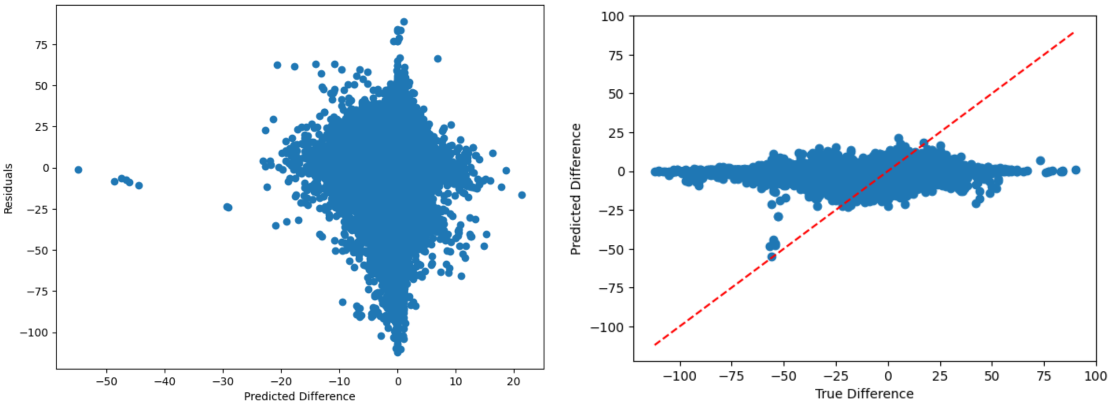
Figure 4.9: Polynomial Regression (PR)

If additional second degree polynomial terms are allowed to the ULR model, the performance of the model can be slightly improved. The resulting **Polynomial Regression** (PR) gives a RMSE of 7.5774, a MAE of 3.7888 and a R^2 of 0.01019. The equation of the PR model is the following:

$$y = -0.20 + 770.15 * X_{10} + 428.81 * X_{10}^2$$

Figure 4.9b shows that the model is able to predict higher values for the predicted difference when the true difference is high. This results in a more accurate residual plot. Figure 4.9a shows that when the predicted difference is higher, the residuals get smaller with few outliers. However, when the predicted difference is lower, the residuals tend to be more negative.

The simple CNN model achieves a RMSE of 7.4899 and a MAE of 3.6018. Figure 4.10a shows that the highest residuals appear in the range of the predicted difference around 0. If the model predicts higher values, the residuals tend to be in the negative range and are relatively accurate. The predicted differences have a larger scatter and are not centered as strongly around 0 as with other models. This is also visible in Figure 4.10b, which shows the true and predicted values.

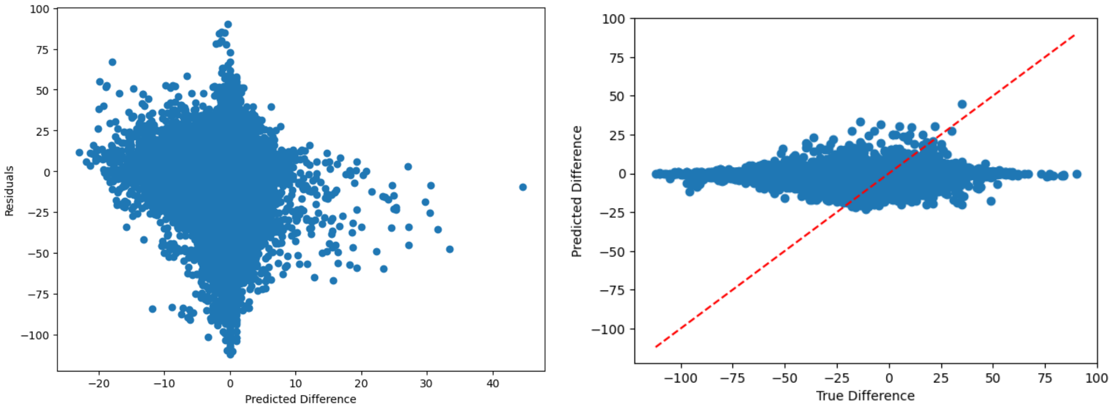


(a) Predicted Difference vs Residuals (b) True Difference vs Predicted Difference

Figure 4.10: CNN

The proposed **CnnTrans** model, which combines four CNN layers with a transformer layer, achieves a RMSE of 7.5080 and a MAE of 3.6052. Figure 4.10 shows the comparison between the predicted values and the true values, as well as the residuals. Again, the largest deviations are in the direction of the extremes. One advantage of this model is that the predictions are no longer so strongly centered around the value 0 and the tendency toward larger values is stronger than in the previous models. When comparing the CNN with the CnnTrans model, it can be seen that the CNN tends to predict in the negative range, while the CnnTrans tends to predict in the positive range.

Table 4.2 shows the comparison of the true values with the predicted values. It also outlines the difference (bias) between them as a Five-Number summary of the models: the Multiple Linear Regression (MLR), the adapted Multiple Linear Regression



(a) Predicted Difference vs Residuals (b) True Difference vs Predicted Difference

Figure 4.11: CnnTrans

(MLR2), the Univariate Linear Regression (ULR), the Polynomial Regression (PR), the CNN and the CnnTrans model. The so-called bias in the following refers to the absolute difference between the true values (y_{true}) and the predicted values (y_{pred}).

		MLR		MLR2		ULR	
	y_{true}	y_{pred}	bias	y_{pred}	bias	y_{pred}	bias
count	598578	598578	598578	598578	598578	598578	598578
mean	-0.740	-0.067	3.773	-0.067	3.759	-0.062	3.761
std	7.531	0.610	6.569	0.539	6.575	0.559	6.575
min	-112.000	-7.369	0.000	-1.111	0.001	-1.159	0.002
25%	-2.000	-0.484	0.766	-0.450	0.761	-0.444	0.759
50%	0.000	-0.135	1.707	-0.134	1.685	-0.117	1.687
75%	2.000	0.246	3.610	0.196	3.580	0.211	3.581
max	90.000	3.709	111.183	3.027	111.248	3.159	111.228

		PR		CNN		CnnTrans	
	y_{true}	y_{pred}	bias	y_{pred}	bias	y_{pred}	bias
count	598578	598578	598578	598578	598578	598578	598578
mean	-0.740	-0.080	3.789	0.004	3.602	-0.010	3.605
std	7.531	0.728	6.562	1.065	6.567	1.084	6.586
min	-112.000	-0.548	0.001	-54.808	1.287	-22.904	0.000
25%	-2.000	-0.494	0.608	0.001	0.969	0.000	0.986
50%	0.000	-0.311	1.548	0.001	1.466	0.000	1.447
75%	2.000	0.002	3.546	0.200	3.194	0.001	3.136
max	90.000	8.629	111.452	21.338	112.033	44.611	112.001

Table 4.2: Five-Number-Summary of all models

The true difference between the wearable and the ECG is distributed in a range of -112 to +90 beats. The predicted difference from the two Deep Learning models (CNN and CnnTrans) tend to center around 0 with a mean of 0.004 and -0.010 resp and a range of [-54.8; 21.4] and [-22.9;44.6] resp. Although these values seem to be small, they are still better than those of the regression models, which have a smaller range, namely [-7.4; 3.7] for the MLR, [-1.1; 3.0] for the MLR2, [-1.2; 3.2] for the ULR and [-0.1; 8.6] for the PR. As forementioned before, the CNN model tends to predict values more likely in the negative range while the CnnTrans tends to predict more likely in the positive range. However, all models do not show very good results since an accurate prediction seems to be very difficult for all of them. For this reason, we allow the model a tolerance range of ± 3 beats in which the predictions are seen as correct and accurate. There are some studies which discuss the boundaries for such a tolerance range for HR measurements, but all authors admit, that further research has to be done in that area. Khan et al. [75], the CTA [10], Haynie et al. [62] and Nelson et al. [106] set the boundary to ± 10 beats or less, since their study found that a difference of 10 bpm or less between two measurements is considered to be clinically insignificant. However, we tightened the limits to get more stringent and accurate values. Since the users should be informed of deviations later, the tolerance range should be as small as possible. Table 4.3 shows the results of accurate measurements relative to the predictions of this model.

	within range ± 3
MLR	70.14%
MLR2	70.55%
ULR	70.51%
PR	69.74%
CNN	72.09%
CnnTrans	72.09%

Table 4.3: Amount of accurate predictions (± 3) for each model

The CNN and the CnnTrans have the best score by predicting 72.09% correct. However the CNN predicts 431494 out of 598578 cases correct, while the CnnTrans predicts 431486 out of 598578 cases correct within the range. All regression models seem to have more problems in predicting correct values even with a tolerance range of 3 beats.

When dividing the measurements of all models into the four predefined intensity ranges, it can be seen that the intensity ranges have an influence of the size of the bias. Table 4.4 shows the relative number of cases where the bias occur. When using the MLR, there are 17 windows which have a larger bias than 100. This means that the true difference and the predicted difference of the model differ more than 100. And out of those 17 windows, 76.47% occur in the minor intensity range.

	bias >	high	moderate	light	minor	absolute number
MLR	100	0.00	5.88%	17.65%	76.47%	17
	90	62.35%	1.18%	3.53%	32.94%	85
	80	67.34%	0.50%	3.52%	28.64%	199
	70	81.49%	1.21%	2.01%	15.29%	497
	60	70.29%	17.01%	1.54%	11.17%	976
	50	59.99%	25.61%	3.15%	11.25%	1812
MLR2	100	0.00	5.88%	17.65%	76.47%	17
	90	63.74%	1.10%	3.30%	31.87%	91
	80	67.16%	0.49%	3.43%	28.92%	204
	70	81.36%	1.20%	2.00%	15.43%	499
	60	70.52%	16.68%	1.54%	11.26%	977
	50	59.63%	26.22%	3.05%	11.10%	1838
ULR	100	0.00%	5.88%	17.65%	76.47%	17
	90	63.74%	1.10%	3.30%	31.87%	91
	80	67.16%	0.49%	3.43%	28.92%	204
	70	81.33%	1.20%	2.01%	15.46%	498
	60	70.55%	16.66%	1.53%	11.25%	978
	50	59.66%	26.24%	3.05%	11.05%	1837
PR	100	0.00%	5.88%	17.65%	76.47%	17
	90	60.71%	1.19%	3.57%	34.52%	84
	80	66.16%	0.51%	3.54%	29.80%	198
	70	81.33%	1.20%	2.01%	15.46%	498
	60	70.91%	16.46%	1.52%	11.11%	990
	50	60.60%	25.32%	3.01%	11.07%	1825
CNN	100	5.26%	5.26%	15.79%	73.68%	19
	90	54.43%	1.27%	3.80%	40.51%	79
	80	67.16%	0.50%	3.48%	28.86%	201
	70	80.89%	1.22%	2.24%	15.65%	492
	60	68.87%	17.53%	1.55%	12.06%	970
	50	58.34%	25.97%	3.34%	12.35%	1798
CnnTrans	100	5.26%	5.26%	15.79%	73.68%	19
	90	55.56%	1.23%	3.70%	39.51%	81
	80	67.00%	0.50%	3.50%	29.00%	200
	70	81.14%	1.22%	2.03%	15.62%	493
	60	69.83%	16.70%	1.57%	11.90%	958
	50	58.21%	25.86%	3.22%	12.71%	1802

Table 4.4: Percentage of predictions that differ more than 'bias'

All models have the greatest difficulty in making good predictions in the high range. Exceptions to this are the cases where the largest deviations (>100 beats) occur. These occur mainly in the minor intensity range. The CNN model shows the smallest number of cases (1798) where the predictions deviate by more than 50 beats.

Even with the tolerance range of ±3 beats, the accuracy of the predictions are not satisfying for a sufficient valid prediction. Therefore, the intention now is to at least elicit the times when the HR measurement of the wearable deviates strongly in order to inform the user of this strong deviation. Since the two Deep Learning models outperformed the Linear Regression models, further investigations will be done using those two models.

4.4.1 Development of a Warning System

Although the predictions were not accurate enough to offer a real-time "correction" of HR streams, wider bounderies provide greater accuracy and thus the opportunity to develop a system that can detect large deviations. Thus, the next step is to develop a system that alerts users when there is a significant difference between the measured heart rate and the actual heart rate, indicating a possible measurement error. Therefore it is needed to define a threshold, which indicates the range from which a PPG measurement is to be regarded as not acceptable. From the previous literature research, a recommendation of ± 10 beats as a validation limit has emerged [10, 62, 75, 106]. Thus, all PPG windows that have a true difference from the ECG of more than 10 beats are considered as 'inacceptable', as shown in Figure 4.12. For the remainder of this thesis, the absolute true difference will be referred to as y_{true}, while the predicted difference will be referred to as y_{pred}.

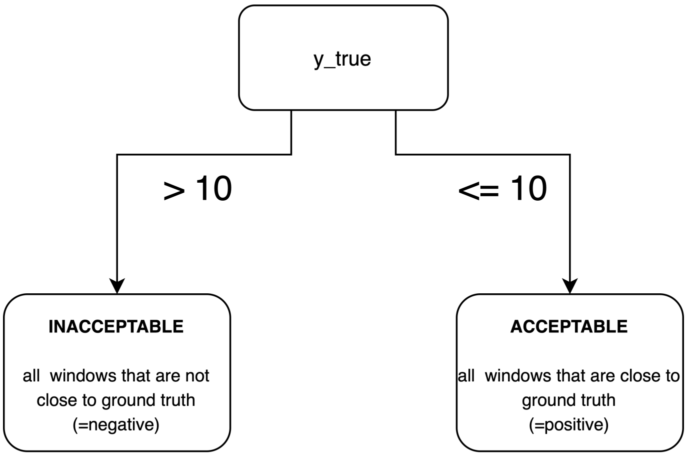

y_true
> 10
<= 10
INACCEPTABLE
all windows that are not close to ground truth (=negative)
ACCEPTABLE
all windows that are close to ground truth (=positive)

Figure 4.12: Sketch of Classification of PPG window

Now a threshold value (τ) is introduced that applies to the magnitude of the predicted value. The larger the model predicts the value, the more likely the model is to believe that the measurement from the wearable on the ground truth is off at that moment.

- If the predicted value $y_{pred} \leq \tau$, the wearable assumes that the window does not deviate much from the ground truth and so the PPG measurement is satisfying.
- If the predicted value $y_{pred} > \tau$, the wearable assumes that the window deviates strongly from the ground truth and so the PPG measurement is not satisfying.

Figure 4.13 shows the structure of the confusion matrix, while Table 4.5 depicts the confusion matrix itself. All windows that are classified as an unacceptable measurement due to an elevated y_{true} are considered as non-satisfying. The performance of the model depends on how many unacceptable cases the model correctly detects as not satisfying (= True Negative) without incorrectly classifying too many acceptable cases as not satisfying (= False Negative).

	$y_{true} \leq 10$ acceptable	$y_{true} > 10$ inacceptable
$y_{pred} \leq \tau$ satisfying	True Positive TP	False Positive FP
$y_{pred} > \tau$ not satisfying	False Negative FN	True Negative TN

Table 4.5: Confusion Matrix

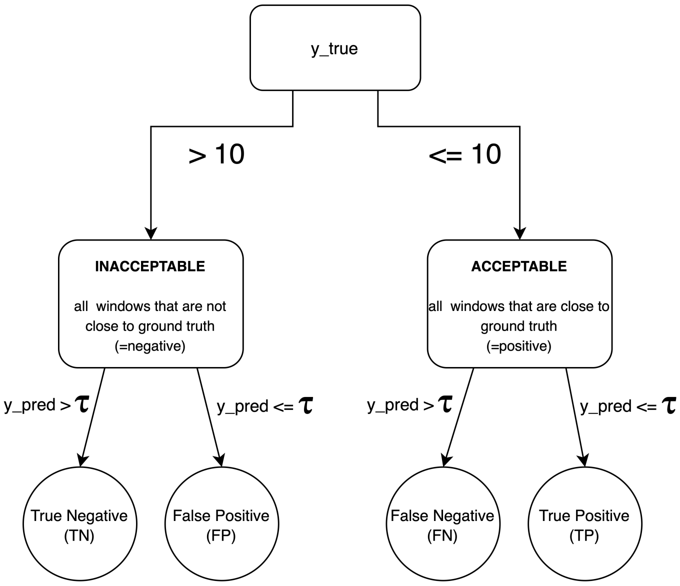


Figure 4.13: Sketch of Confusion Matrix

After preliminary tests, ranges of possible values for τ = [10, 20, 25, 30, 31, 32, 33, 34, 35, 40] are defined. The optimal values can be determined based on the accuracy of the Performance (Perf) capture 4.2.

$$\text{Performance (Perf)} = \frac{TN}{TN + FN} * 100 \tag{4.2}$$

	CNN					CnnTrans				
τ	TN	FP	FN	TP	Perf	TN	FP	FN	TP	Perf
10	348	50998	255	546977	57.71	500	50846	314	546918	61.43
20	16	51330	3	547229	84.21	19	51327	18	547214	51.35
25	8	51338	0	547232	100.00	6	51340	5	547227	54.55
30	6	51340	0	547232	100.00	3	51343	2	547230	60.00
31	6	51340	0	547232	100.00	2	51344	1	547231	66.67
32	6	51340	0	547232	100.00	2	51344	0	547232	100.00
33	6	51340	0	547232	100.00	2	51344	0	547232	100.00
34	6	51340	0	547232	100.00	1	51345	0	547232	100.00
35	6	51340	0	547232	100.00	1	51345	0	547232	100.00
40	6	51340	0	547232	100.00	1	51345	0	547232	100.00
45	5	51341	0	547232	100.00	0	51346	0	547232	-
50	1	51345	0	547232	100.00	0	51346	0	547232	-

Table 4.6: Perf of CNN and CnnTrans

Table 4.6 shows the results of this investigation when applying this on the results of the CNN and CnnTrans. The CNN model is able to detect more True Negative (TN) cases even with a small τ. The performance of only 25 gives a 100% certainty that the current window deviates from the ground truth by more than 10 beats and is detected as inacceptable. In comparison, the CnnTrans needs a threshold of 32 or more to get a 100% certainty. With a threshold of 25, the probability of detecting a true measurement error is only about 55%.

Linear regression models were not able to predict values higher than 10, which is a limitation for developing a warning system. This is because warning systems need to be able to predict values that are outside of the normal range in order to provide early warning of potential problems. In contrast, the CNN model was able to predict higher values with good performance. This suggests that the CNN model is the best choice for developing a warning system.

By setting a threshold τ, it depends on the size of τ at which point the warning system steps in and classifies the measurement as not satisfactory. Figure 4.14 shows the structure of the process used to develop the warning system. A soon as y_{pred} exceeds τ_1, the user gets a 'pre-warning', which states that this measurement has a high potential to be unacceptable. As soon as y_{pred} exceeds τ_2, the user gets a warning which gives a high probability of an error.

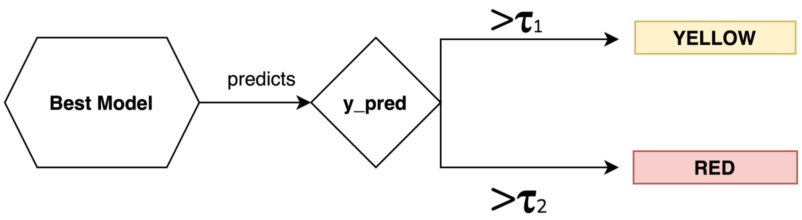


Figure 4.14: Sketch of warning system

According to the previous investigations, the thresholds 20 and 25 show the most plausible values for being taken as thresholds when using the CNN model. If the value exceeds 20, the model can say with a confidence of 84% that this measurement is not satisfactory. If the value exceeds 25, there is a 100% confindence the value is inaccurate.

4.4.2 Evaluation of the Warning System

The CNN model was originally trained on a dataset of heart rate data collected from a specific wearable, namely Garmin venu2s. However, the goal of the warning system was to be applicable to different wearables. To test the generalizability of the model, it was re-trained on a new dataset of heart rate data collected from the same study, but using a different wearable device (Polar Verity Sense). The new dataset was recorded at the same frequency (1 Hz) and used the same protocol as the original dataset. The data was also preprocessed and synchronized with the gold standard in the same way.

The CNN model was then applied to the new dataset and the same calculations were performed to evaluate the different thresholds. The results of these calculations are shown in Table 4.7. The results show that the CNN model is able to generalize to different wearable devices and still achieve high accuracy. The new data set (Polar) shows similar performance values to those of the original data set (Garmin).

Two thresholds, $\tau_1 = 20$ and $\tau_2 = 34$, were found to work reasonably well for both models. The CNN model correctly detects about 84% of unacceptable measurements in the original dataset when the first threshold, τ_1, was set to 20. This threshold can be considered a pre-warning level. For the new dataset, the model correctly detects about 86% of the unacceptable measurements at the same threshold value. When the second threshold, τ_2, was set to 34, the model detected 100% of all inacceptable measurements in both data sets. This threshold value can be seen as a reliable warning for measures that are substantially off.

	Polar					Garmin
τ	TN	FP	FN	TP	Perf	Perf
10	419	34830	277	468400	60.20	57.71
20	67	35182	11	468666	85.90	84.21
25	29	35220	2	468675	93.55	100.00
30	11	35238	1	468676	91.67	100.00
31	9	35240	1	468676	90.00	100.00
32	5	35244	1	468676	83.33	100.00
33	3	35246	1	468676	75.00	100.00
34	3	35246	0	468677	100.00	100.00
35	2	35247	0	468677	100.00	100.00
40	1	35248	0	468677	100.00	100.00
45	0	35249	0	468677	-	100.00
50	0	35249	0	468677	-	100.00

Table 4.7: Performance of CNN on Polar Data in comparison to the original Garmin Data

5 Discussion

The first part of this thesis examined the **validity of HR measurements made with wearable devices**, specifically addressing the question when large measurement errors are likely to occur. When talking about the term "validity", it must be emphasized that there is no official framework defining when a device is to be considered valid or not. Testing standards are also not transparent and remain unknown to the consumer. Each manufacturer decides for itself whether the validation of a wearable meets medical certifications. This leads to a large heterogeneity of validation protocols. According to Mühlen et al [102], the assessment of validity should be performed by independent institutions. The constantly growing number of new wearables makes it almost impossible for scientific institutions to keep up with the latest developments. This is also criticized by Carrier et al. [27] in their paper. Therefore, Mühlen et al. [102] propose a common framework for assessing validity that can be used by both manufacturers and research institutions. Studies to determine validity should evaluate the instrument using a precise measurement criterion in a relevant sample and under conditions that reflect expected use in the real-world. The evaluation should be properly recorded and described in an understandable manner. Back in 2018, a preliminary framework was presented by the Consumer Technology Association [124], however it did not include recommendations for long-term HR monitoring under field conditions. Scientific evidence for these proposed guidelines was also not provided. In the future, as of May 2025, all wearables must comply with medical device regulations, such as CE marking in Europe.

Another important challenge in assessing the validity of wearables is that studies that investigate them employ different study designs and methods, reducing replicability and comparability. Many studies use an ECG measurement as a comparative method or ground truth, but some studies resort to chest straps or pulse oximeters instead [23, 41, 45, 147]. These methods may again have some error and are thus only suboptimal methods of comparison, and the results of these studies should be interpreted with caution [106].

Both a literature review and an analysis of real-world data revealed that wearable devices have the most problems in high-intensity ranges [13, 19, 24, 40, 43, 58, 68, 125, 153, 161]. Although this difference regarding exercise intensity could not be found in all studies [41, 136, 145, 170], the investigation of real-world data confirmed this. This findings raise the question of whether the increased HR and the resulting faster changes in blood volume in the vessels, are the reason for those problems or whether there are other reasons for the deviations. One hypothesis is that the blood volume changes in the vessels occur too fast for the wearables to capture. However the measurement problems could also occur because of the increased movements the subject makes during the measurement. Nelson et al. [106] pointed out that activity intensity may be less important for accuracy than the amount of irregular wrist movements during physical activity. These movements tend to be higher during intense physical activity and may cause the sensors to slip or lose contact with the skin, resulting in motion artifacts. Various authors considered this possibility as well and also believe that these motion artifacts are the biggest problems for decreased accuracy. Both Navalta et al [105], Schäck et al [133], Essalat et al [46], and Zhang et al [176] suggest that higher intensity means more motion. They see the increased arm movements as the main factor for motion artifacts, which then lead to erroneous measurements. Few studies have compared the accuracy of wearable devices for measuring HR during high-intensity physical activity with a lot of arm movements (e.g., running) to those with less arm movements (e.g., cycling). Some devices performed better during running, while others performed better during cycling [113]. Reddy et al. [125] discovered that validity decreases at high intensity ranges with even a lack of wrist movement. However, motion artifacts can also be caused by abnormal blood pressure changes if the wearer has hypertension [46].

Ambient light, in addition to motion artifacts, could also affect the measurement. While many studies conduct their investigation in a laboratory-like environment [23, 41, 48, 58, 68, 125, 145, 147, 153, 160] the influence of ambient light may be higher when training outdoors. Recent research has challenged the validity of laboratory studies, as they are conducted in controlled and artificial environments that may not reflect the real-world experiences of participants. Consumers use wearables in natural environments where physiological and psychological situations may occur that may not be replicated and captured in the laboratory [105, 106, 132]. In addition to motion artifacts and ambient light, other factors such as ambient temperature, melatonin concentration, skin pigmentation, and body hair may also play a significant role in the real-world measurement accuracy of wearables [44, 48, 58, 102, 105, 151, 154, 160, 176].

French et al. [51] do not address measurement with fitness wearables due to the lack of valid research. They also describe a lack of agreement in the literature at both submaximal and maximal exercise intensities, which makes data interpretation more difficult. They attribute this to the many dynamic and multidimensional physiological conditions that can affect wearable measurements, such as rapid changes in intensity, body position, the psychological and physical environment, individual training goals and the sport being performed [22, 51]. Authors therefore recommend being aware of the limitations of wearables and only use them in conjunction with additional aids such as a chest strap or accelerometry [24]. There is some research showing that the accuracy of wearables is better when combined with accelerometry [74, 112, 160].

Overall, it is important to view high-range HR measurements with caution. As mentioned earlier, prolonged high HR is a risk factor for CVD. Physical activity can reduce HR and therefore help to prevent CVD and other chronic diseases. Accurate and valid HR measurements are needed to individualize physical training appropriately [160]. This is true in both the medical and health fields, as well as in sports. One of the biggest challenges in exercise design is to find the right balance between exertion and recovery. One way to do this is to measure the internal responses to exercise, such as HR. Heart rate can be used to calculate relative intensity ranges, which can then be used to plan appropriate training. High-intensity exercise requires longer rests to allow for full cardio-autonomic recovery. Low-intensity exercise requires up to 24 hours of rest, while moderate-intensity exercise requires 24-48 hours of rest. As fitness improves, the stress-recovery adaptation cycle becomes shorter. Heart rate monitoring is a valuable tool for understanding and improving individual fitness. It can be used to control intensity and effectively design and vary a training plan. Monitoring can improve fitness, recovery time and overall performance.[51, 100, 146]. Valid and accurate measurements are therefore important to give correct advices and to not overcharge or underchallenge an individual.

Most wearables use proprietary algorithms to convert PPG signals into HR measurements and estimates. These algorithms are not transparent to outsiders and are regularly updated. It is also problematic that most authors do not specify the firmware version of the wearable device used in their studies. This can lead to discrepancies between studies that use the same device but different firmware versions, as the firmware can affect the accuracy of the device's measurements. Lack of reproducibility can make it difficult to draw reliable conclusions from research on wearable devices. Additionally, the rapid pace of development in the wearables industry means that devices used in studies may quickly become outdated. This is because wearable devices are constantly being updated with new features and algorithms, which can improve their accuracy. As a result, research on wearable devices may always lag behind the latest technological developments [106, 167]. One way to address this issue is to conduct research using a different design. For example, researchers could conduct longitudinal studies that track the same participants over time as they use different wearable devices. Although people also become more different from themselves over time, this option would still be the most comparable. This would allow researchers to compare the accuracy of different devices and firmware versions over time. Another way to address this issue is to develop standardized protocols for testing the accuracy of wearable devices. This would ensure that all studies are conducted using the same methods, which would make it easier to compare results from different studies. Ultimately, the challenge of keeping up with the rapid pace of technological advancement in the wearables industry is a significant one. However, by using different research designs and developing standardized protocols, researchers can help to ensure that research on wearable devices is as reliable and up-to-date as possible.

Classification versus (VS) Regression

A key question at the beginning of the thesis was how to define a measurement error. One possibility was to manually classify windows as "on" or "off" in advance and use this data as a supervised method to train the model. "On" means that the measurement of the wearable is correct, while "off" corresponds to an incorrect measurement. However, this method is time-consuming and not transparent, since the classification is done manually and arbitrary. Possible options could be to compare the 10 second window regarding the RMSE and classify to a certain arbitrary threshold. Another approach could be to compare the absolute differences between the measurements and set a threshold for the distance. Both approaches yield the same problem: setting the threshold is arbitrary and, therefore, objective. Additionally, the classes should be even, which is hard to find methods that achieves this without losing too much data. Therefore the approach of this thesis is a prediction method. The idea is to first get a prediction from the model, and then use a threshold to determine whether the current measurement should be classified as "on" or "off". This led to the development of a warning system. If the model's prediction is high and exceeds a certain threshold, it can be assumed that the measurement is most likely wrong. The warning system distinguishes between two different warning levels, depending on the level and probability of the deviation.

Model Selection

Finding a suitable and 'best' model was very time-consuming and computationally intensive. The model had to be accurate enough to predict HR, but it should also generalize well to external data and other wearables. A more complex model can represent the underlying relationship more accurately, but it is also more likely to overfit the training data. Overfitting occurs when the model learns the noise in the training data instead of the underlying relationship. This can lead to poor performance on new data. The goal is to find a model with low variance and low bias. Variance describes the error caused by overfitting, while bias describes the error introduced by the assumptions of the algorithm chosen to build the model. A model with low variance will have a small error on the training data, but it may have a large error on new data. A model with low bias will have a large error on the training data, but it may have a small error on new data. The model chosen in this study had slightly reduced performance on the training set, but it had the best performance on subsequent test sets. This suggests that the model has low variance and low bias, and is therefore generalizable to new data [51]. However, neural networks are known to be data-dependent. This means that their performance can vary depending on the training data used. Additionally, neural networks are often prone to overfitting. This can lead to unstable predictions and is therefore still questionable in real-world applications [16]. For that, the approach was performed on a data set with a different wearable. Nevertheless, there is a possibility that the thresholds may have been overfitted to these two data sets. Further research with additional data sets is needed to elicit better and more generealizable thresholds.

For **data preparation**, the HR measurements of the ECG and the wearable were used. All missing values were removed from the two measurements, and obvious errors and outliers were removed from the ECG. The ECG was used as the ground truth, therefore, it was important to have it as clear and error-free as possible. One option would have been to fill in the errors with other data, such as the mean or median. However, this was not done because it would have introduced fictional data and altered the ground truth. The PPG measurement was compared one-to-one with the ECG, so it was important to keep the PPG measurement as unaltered as possible. Outliers in the PPG measurement were not removed so as not to embellish the measurements and to take them exactly as they occur in real life.

The data was then split into windows. The size of the window is an important parameter that can affect the performance of many data mining tasks, such as classification, clustering, anomaly detection and time series prediction. The optimal window size can be difficult to determine, and is often specified by experts or based on experience. If the window size is too small, important patterns in the data may be missed. If the window size is too large, important patterns may be distorted or averaged out. In this work, different window sizes were suggested by experts. The window sizes with the best performance were used for further processing.

Deep Learning Approach

In this work, two deep learning models were compared: a CNN model and a CnnTrans model. The CnnTrans model was expected to outperform the CNN model due to the advantages of transformers, such as their ability to capture long-range dependencies, such as fluctuations and dynamics over time, and their scalability. However, the CNN model showed slightly better performance. The results of this study suggest that the CNN model is a better choice for detecting measurement errors in this particular dataset. It is important to note that the performance of the models may vary depending on the dataset and the specific application. A possible reason why the CNN model outperformed the CnnTrans model could be the size of the dataset. The dataset may not have been large enough to fully exploit the advantages of transformers. Also, the CNN model may have been better at capturing the specific features of the dataset that were relevant to detecting measurement errors. What is more is that the CNN model may have been simpler and easier to train, potentially resulted in a more accurate model. Further research is needed to determine the best model for detecting measurement errors in different datasets and applications.

In addition to comparing the results of the models, the performance of the calculation with new values was also evaluated. This is important because the warning system is intended to inform the user in a timely manner when measurement errors occur. To accomplish this, the test data was passed to the model one at a time and the time it took for the model to calculate the appropriate output was calculated. For the simulation, 10,000 values (= seconds) were used. Again, the CNN model showed slightly better performance, with a mean of 0.1038 seconds. The CnnTrans model has a mean of 0.1183 seconds. The remaining values for analysis can be seen in Table 5.1.

Computation Time	CNN	CnnTrans
Minimum	0.0622	0.0709
First Quartile	0.0869	0.0935
Median	0.0961	0.1072
Third Quartile	0.1132	0.1319
Maximum	0.9658	1.6532
Mean	0.1038	0.1183
Standard Deviation	0.0305	0.0428

Table 5.1: Five Number Summary of Computation Time of CNN and CnnTrans in seconds

Purpose and Limitation of the used approach

The increased interest of research towards wearables and the resulting rapid growth of data comes the idea of analyzing these data with Machine Learning and Deep Learning to generate further knowledge. Using these techniques, the goal is to adapt and improve HR estimates from the wearables. While Alharbi et al [5] writes that HR data are considered non-stationary, meaning that they are constantly changing and therefore cannot be predicted or modeled some research has already been done and algorithms created to classify or predict HR data. However, these studies focused on either processing the raw PPG signal with their own algorithms to estimate HR, attempting to classify heart disease using ECG measurements, or using higher frequency PPG signals [18, 77, 104, 108, 139, 176]. The goal of this work was to validate existing HR measurements while keeping the calculation as simple as possible. This was done in order to require little preprocessing time and battery power, while still generating accurate information with the simple model. The intended outcome was to provide immediate feedback to the end user in the event of significant deviations or measurement errors. This approach should be easy to implement on different wearables, making it device-independent and generalizable to other applications.

The disadvantage of this approach is that it does not consider most of the information that the PPG signal provides. The PPG signal is a rich source of information about HR, however, it only uses a small portion of this information. This approach is based on an algorithm developed by the device manufacturer and is not visible to outsiders. This means that it is not possible to know exactly how the algorithm works or how it is making its predictions. It also draws on data collected at a frequency of 1 Hz. This means that the data is only sampled once per second. If the data were collected at a higher frequency, more information could be extracted from the temporal course of the signals. However, these options are omitted because the model should be kept as simple and as general as possible. The goal of this approach is to develop a model that can be used with a variety of wearable devices. If the model were too complex, it would not be able to run on all devices. This approach deals with a real-world problem and with the output displayed by the wearable. The PPG signal is not accessible on the wearable because it is already processed in advance.

Another important point is that the model shows a large number of False Positive (FP) cases for both data sets (Garmin and Polar). At the two thresholds ($\tau_1 = 1$, $\tau_2 = 2$) leads to a very high sensitivity and a very low specificity. The formulas for sensitivity and specificity are as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

The exact values can be found in the table 5.2. The **sensitivity** indicates how many of the false measurements are actually detected by the model. A high sensitivity means a high certainty of detecting false measurements. The **specificity** indicates the probability that correct measurements are actually recognized as correct by the model. It is a measure of how high the proportion of correct measurements is, which are also recognized as correct.

data set	threshold	sensitivity	specificity
Garmin	20	0.9999	0.0003
	34	1	0.0001
Polar	20	0.9999	0.0019
	34	1	0.0001

Table 5.2: Sensitivity and Specificity of Data Sets

A high sensitivity and a low specificity means that the model classifies many measurements as incorrect, but they are actually correct. In the proposed CNN model, this low specificity is a major weakness. This should be revised and developed further, since there are a large number of false positives.

In addition, it must be mentioned that the group of subjects consists of healthy patients, but also 11 patients who took HR influencing drugs. The HR records of these patients could influence the measurements and make it difficult for the model to make appropriate predictions. This must be taken into account when analyzing the results. On the other hand, CVD patients are not excluded from the possibility of using this model in the future.

6 Future work

The presented CNN model has the potential to provide the foundations for a warning system to inform users of very large deviations from the gold standard. The model works well on the two datasets presented, achieving up to 100% confidence in identifying inadequate measurements. However, the model is not yet perfect, as the accuracy of the predictions is not always successful.

Follow-up research should focus on refining and improving the input data, such as frequency and format, to make better predictions thereby predicting measurement inaccuracies. It may be possible to work directly with the PPG signal instead of using the HR values issued by the wearable device. This would retain more information from the signal and allow the data to be processed at a higher frequency.

In addition, an important point for future work is to address how the newly obtained information and results from the model are communicated to the end user. This area in HDI has a lot of potential, but still requires a lot of research: what information is important to the end user? What kind of communication is sufficient, and how much is too much? How do they react to the thresholds?

Depending on the model, it is possible to give warnings to the end user promptly or with a certain time delay. There would also be the possibility to count the occurrence of a measurement error and return the total sum to the end user at the end of the training, along with the summary. At the same time, the warnings could be announced via vibration or sound immediately upon occurrence. Further research is needed in this area.

7 Conclusion

Wearables have become increasingly popular in recent years, as they offer a convenient and non-invasive way to monitor health metrics such as HR. However, the accuracy of HR measurements from wearable devices can be affected by a number of factors. Wearables show the most difficulties and inaccuracies especially in the high intensity range. Although the manufacturers and companies of the wearables are constantly trying to improve the HR estimates, many improvements are still needed.

In this thesis, two Deep Learning models and four Linear Regression models were compared. One of the Deep Learning models based on a CNN architecture outperformed the other models. The CNN model can be used to detect the large deviations and thus the occurrence of inacceptable measurements. It is trained on a dataset of real-world HR measurements, and it can be used to inform and warn users about large deviations from the gold standard. The model works well on two datasets presented, achieving up to 100% reliability in detecting insufficient measurements.

List of Figures

2.1 The P-QRS-T complex of a normal ECG wave. Adapted from Zhang ([174] 6
2.2 The PPG signal compared to the corresponding ECG recording of the HR. Adapted from Allen et al. [6]. 7
2.3 Examples of different types of MA on index finger. Adapted from Allen et al. [6] 8
2.4 Sketch of a Perceptron 12
2.5 Comparison of the function of a sigmoid neuron (left) versus the function of a perceptron (right) 13
2.6 Basic architecture of a Artificial Neural Network 14
2.7 Architecture of a simple CNN 18
2.8 Each Convolution Layer involves Convolution, Non-linearity and Pooling. 20

3.1 Recording relative number of errors of wearable (PPG) 27
3.2 Boxplot of error metrics and Pearson's Correlation Coefficient between ECG and wearable separated into different intensity ranges 30
3.3 Total values of error metrics of different intensity ranges 31
3.4 Total values in different intensity ranges 31
3.5 Bland Altmann Plots of separate ranges 32
3.6 Comparison of btb versus avg of different error metrics over separate ranges 34

4.1 Data preprocessing for supervised learning 41
4.2 Preprocessing for Linear Regression model 42
4.3 Architecture of Deep Learning model 43
4.5 Multiple Linear Regression (MLR) 45
4.6 Independent variable X_1 to true difference in test data set 46
4.7 Adapted Multiple Linear Regression (MLR2) 47
4.8 Univariate Linear Regression (ULR) 48
4.9 Polynomial Regression (PR) 48
4.10 CNN 49
4.11 CnnTrans 50
4.12 Sketch of Classification of PPG window 53
4.13 Sketch of Confusion Matrix 54
4.14 Sketch of warning system 56

7.1 The remaining variables of the Multiple Linear Regression (MLR) model I. XXIII
7.2 The remaining variables of the Multiple Linear Regression (MLR) model II. XXIV

List of Tables

3.1 Summary of measurement of HR data of ECG and wearable HR mean ... mean of all HR measurements ± standard deviation HR max ... maximal HR value HR min ... minimum HR value 25
3.2 Errors in ECG and wearable measurements: mean ± standard deviation 26
3.3 Recording errors of wearable measurement in different intensity ranges 27
3.4 Comparison of data before and after the cleaning process. 28
3.5 Mean ± standard deviation of error metrics of all subjects 29
3.6 Total Values of Error Metrics 29
3.7 Total values of Pearson's Correlation Coefficient (corr) and Matching Beats (match) 31
3.8 Bland Altmann Plot values of different ranges 32
3.9 Comparison of beat-to-beat (btb) vs 10 second average (avg) 33

4.1 Number of windows in training and test set 41
4.2 Five-Number-Summary of all models 50
4.3 Amount of accurate predictions (± 3) for each model 51
4.4 Percentage of predictions that differ more than 'bias' 52
4.5 Confusion Matrix 54
4.6 Perf of CNN and CnnTrans 55
4.7 Performance of CNN on Polar Data in comparison to the original Garmin Data 57

5.1 Five Number Summary of Computation Time of CNN and CnnTrans in seconds 63
5.2 Sensitivity and Specificity of Data Sets 64

List of Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
ANS	Autonomic Nervous System
avg	Average Window
BP	Blood Pressure
bpm	Beats per Minute
btb	Beat-to-Beat
CNN	Convolutional Neural Network
CnnTrans	CnnTrans
Corr	Pearson Correlation Coefficient
corr	Pearson Correlation Coefficient
CVD	Cardiovascular Disease
DBN	Deep Belief Network
ECG	Elektrocardiogram
e.g.	exempli gratia
etc	et cetera
FCN	Fully Conventional Network
FFT	Fourier Transform
FN	False Negative
FP	False Positive
GAN	Generative Adversarial Network
GBRT	Gradient Boosted Regression Tree
GRU	Gated Recurrent Unit
HCI	Human-Computer Interaction
HR	Heart Rate
HRmax	Maximum Heart Rate
HRmin	Minimum Heart Rate
HDI	Human-Data Interaction
Hz	Hertz
i.e.	id est
KNN	K-Nearest Neighor
LED	Light-Emitting Diode
LSTM	Long Short Term Memory Network
MA	Motion Artifact

match	Matching Beat
max	maximum
MAE	Mean Absolute Error
MDE	Mean Directional Error
min	minimum
MLP	Multilayer Perceptron
MLR	Multiple Linear Regression
MLR2	adapted Multiple Linear Regression
Perf	Performance
PPG	Photoplethysmography
PR	Polynomial Regression
RBM	Restricted Boltzmann Machine
ReLU	Recitifed Linear Unit
resp	respectively
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
std	standard deviation
SNR	Signal-to-Noise Ratio
SVM	Support Vector Machine
SVR	Support Vector Regression
TN	True Negative
TP	True Positive
ULR	Univariate Linear Regression
VIF	Variance Inflation Factor
vs	versus
WHO	World Health Organization

Bibliography

[1] O. Abdeljaber, O. Avci, S. Kiranyaz, M. Gabbouj, and D.J. Inman. Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks. Journal of Sound and Vibration, 388:154–170, 2017.

[2] J. Achten and A.E. Jeukendrup. Heart rate monitoring: Applications and limitations. Sports Medicine, 33(7):517–538, 2003.

[3] M. Aczon, D. Ledbetter, L. Ho, A. Gunny, A. Flynn, A.M. Gunny, J. Williams, and R. Wetzel. Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks. arxiv, (arXiv:1701.06675), 2017.

[4] R. Al-Hmouz, W. Pedrycz, A. Balamash, and A. Morfeq. Description and classification of granular time series. Soft Comput, 19:1003–101, 2015.

[5] A. Alharbi, W. Alosaimi, R. Sahal, and H. Saleh. Real-time system prediction for heart rate using deep learning and stream processing platforms. Complexity, 2021:1–9, 2021.

[6] J. Allen. Photoplethysmography and its application in clinical physiological measurement. Physiological Measurement, 28(3):R1–R39, 2007.

[7] T. Aoyagi and K. Miyasaka. Pulse oximetry: its invention, contribution to medicine, and future tasks. Anesthesia and analgesia, 94(1 Suppl):S1–3, 2002.

[8] S.O. Arik, H. Jun, and G. Diamos. Fast spectrogram inversion using multi-head convolutional neural networks. IEEE Signal Processing Letters, 26(1):94–98, 2019.

[9] H.H. Asada, P. Shaltis, A. Reisner, S. Rhee, and R.C. Hutchinson. Mobile monitoring with wearable photoplethysmographic biosensors. IEEE Engineering in Medicine and Biology Magazine, 22(3):28–40, 2003.

[10] Consumer Technology Association. Physical activity monitoring for heart rate. `https://standards.cta.tech/kwspub/published_docs/CTA-2065-Preview.pdf`, 12 2018.

[11] Y.Y.M. Aung, D.C.S. Wong, and D.S.W. Ting. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. British Medical Bulletin, 139(1):4–15, 2021.

[12] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arxiv, (1409.0473), 2014.

[13] Y. Bai, P. Hibbing, C. Mantis, and G.J. Welk. Comprehensive evaluation of heart rate-based monitors: Apple watch vs fitbit charge hr. Journal of Sports Science, 36(15), 2018.

[14] Y. Bengio, I. Goodfellow, and A. Couville. Deep Learning. MIT Press, 2016.

[15] B. Bent, B.A. Goldstein, W.A. Kibbe, and J.P. Dunn. Investigating sources of inaccuracy in wearable optical heart rate sensors. NPJ Digital Medicine, 3(18):1–9, 2020.

[16] V.L. Berardi and G.P. Zhang. An empirical investigation of bias and variance in time series forecasting: modeling considerations and error evaluation. IEEE Transactions on Neural Networks, 14(3):668–679, 2003.

[17] M.C. Bishop. Neural Networks for Pattern Recognition. Oxford University Press, USA, 1996.

[18] D. Biswas, L. Everson, M. Liu, M. Panwar, B.E. Verhoef, S. Patki, C.H. Kim, A. Acharyya, C. Van Hoof, M. Konijnenburg, and N. Van Helleputte. Cornet: Deep learning framework for ppg-based heart rate estimation and biometric identification in ambulant environment. IEEE Transactions on Biomedical Circuits and Systems, 13(2):282–291, 2019.

[19] B.D. Boudreaux, E.P. Hebert, D.B. Hollander, B.M. Williams, C.L. Cormier, M.R. Naquin, W.W. Gillan, E.E. Gusew, and R.R. Kraemer. Validity of wearable activity monitors during cycling and resistance exercise. Medicine and Science in Sports and Exercise, 50(3):624–633, 2018.

[20] E. Brophy, W. Muehlhausen, A.F. Smeaton, and T.E. Ward. Optimised convolutional neural networks for heart rate estimation and human activity recognition in wrist worn sensing applications. arXiv, (2004.00505), 2020.

[21] J. Brownlee. Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python. 1.9 edition, 2020.

[22] M. Buchheit. Monitoring training status with hr measures: do all roads lead to rome? Frontiers in Physiology, 5:1–19, 2014.

[23] J. Bunn, E. Wells, J. Manor, and M. Webster. Evaluation of earbud and wristwatch heart rate monitors during aerobic and resistance training. International Journal of Exercise Science, 12(4):374, 2019.

[24] J.W. Bunn, J.A .and Navalta, C.J. Fountaine, and J.D. Reece. Current state of commercial wearable technology in physical activity monitoring 2015-2017. International Journal of Exercise Science, 11(7):503–15, 2018.

[25] F. Cabitza and A. Locoro. Human-data interaction in healthcare, 2016.

[26] L. Cai, K. Janowicz, G. Mai, B. Yan, and R. Zhu. Traffic transformer: Capturing the continuity and periodicity of time series for traffic forecasting. Transactions in GIS, 24(3):736–755, 2020.

[27] B. Carrier, B. Barrios, B. D. Jolley, and J. W. Navalta. Validity and reliability of physiological data in applied settings measured by wearable technology: A rapid systematic review. Technologies, 8(4):70, 2020.

[28] A.V.J Challoner and C.A. Ramsay. A photoelectric plethysmography for the measurement of cutaneous blood flow. Physics in Medicine Biology, 19(3):125–51, 1974.

[29] A. Chandra, L. Tünnermann, T. Löfstedt, and R. Gratz. Transformer-based deep learning for predicting protein properties in the life sciences. eLife, 12, 2023.

[30] N. Charkoudian, J. H. Eisenach, M. J. Joyner, S. K. Roberts, and D. E. Wick. Interactions of plasma osmolality with arterial and central venous pressures in control of sympathetic activity and heart rate in humans. American Journal of Physiology-Heart and Circulatory Physiology, 289(6):H2456–H2460, 2021.

[31] C. Che, P. Zhang, M. Zhu, Y. Qu, and B. Jin. Constrained transformer network for ecg signal processing and arrhythmia classification. BMC Medical Informatics and Decision Making, 21(1):184, 2021.

[32] W. Chen, Z.and Zuo, Q. Hu, and L. Lin. Kernel sparse representation for time series classification. Information Sciences, 292(1):15–26, 2015.

[33] J. Claes, R. Buys, A. Avila, D. Finlay, A. Kennedy, D. Guldenring, W. Budts, and V. Cornelissen. Validity of heart rate measurements by the garmin forerunner 225 at different walking intensities. Journal of Medical Engineering Technology, 41(6):480–485, 2017.

[34] K.J. Coakley and P. Hale. Alignment of noisy signals. IEEE Trans Instrum Meas, 50:141–9, 2001.

[35] A.B. de Luna, V.N. Batchvarov, and M. Malik. The morphology of 1 the electrocardiogram. 2005.

[36] R. Delgado-Gonzales, J. Parak, A. Tarniceriu, P. Renevey, M. Bertschi, and I. Korhonen. Evaluation of accuracy and reliability of pulseon optical heart rate monitoring device. Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pages 430–3, 2015.

[37] J. D. J. Deng and P. Jirutitijaroen. Short-term load forecasting using time series analysis: A case study for singapore. IEEE Conference on Cybernetics and Intelligent Systems, 1:231–236, 2010.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv, (1810.04805), 2018.

[39] World Health Organization Heart Disease. `https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1`, 13.07.3023.

[40] C.J. Dondzila, C.A. Lewis, and J.R. Lopez. Congruent accuracy of wrist-worn activity trackers during controlled and free-living conditions. International Journal of Exercise Science, 11(7):575–84, 2018.

[41] E. Dooley, N. Yarish, and J.B. Bartholomew. Estimating accuracy at exercise intensities: A comparative study of self-monitoring heart rate and physical activity wearable devices. JMIR mHealth and uHealth, 5(3):e34, 2017.

[42] A. Dosovitskiy, L. Beyer, D. Kolesnikov, A.and Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, S. Heigold, G.and Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transfromers for image recognition at scale. International Conference on Learning Representations, 2021.

[43] P. Düking, L. Giessing, M.O. Frenkel, K. Koehler, H.-C. Holmberg, and B. Sperlich. Wrist-worn wearables for monitoring heart rate and energy expenditure

while sitting or performing light-to-vigorous physical activity: Validation study. JMIR mHealth and uHealth, 8(5):e16716, 2020.

[44] O. Dur, C. Rhoades, M. S. Ng, R. Elsayed, R. Van Mourik, and M.D. Majmudar. Design rationale and performance evaluation of the wavelet health wristband: Benchtop validation of a wrist-worn physiological signal recorder. JMIR mHealth and uHealth, 6(10):e11040, 2018.

[45] F. El-Amrawy and M.I. Nounou. Are currently available wearable devices for activity tracking and heart rate monitoring accurate, precise, and medically beneficial? Healthcare Informatics Research, 21(4):315–320, 2015.

[46] M. Essalat, M.B. Mashhadi, and F. Marvasti. Supervised heart rate tracking using wrist-type photoplethysmographic (PPG) signals during physical exercise without simultaneous acceleration signals. IEEE Global Conference on Signal and Information Processing, pages 1166–1170, 2016.

[47] C. Eswaran and R. Logeswaran. An adaptive hybrid algorithm for time series prediction in healthcare. Second International Conference on Computational Intelligence, Modelling and Simulation, pages 21–26, 2010.

[48] M. Etiwy, Z. Akhrass, L. Gillinov, A. Alashi, R. Wang, G. Blackburn, S. M. Gillinov, D. Phelan, A.M. Gillinov, P.L. Houghtaling, H. Javadikasgari, and M.Y. Desai. Accuracy of wearable heart rate monitors in cardiac rehabilitation. Cardiovascular Diagnosis and Therapy, 9(3):262–271, 2019.

[49] H.I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.A. Muller. Deep learning for time series classification: a review. Data Mining and Knowledge Discovery, 33(4):917–963, 2019.

[50] A.A. Flatt and M.R. Esco. Validity of the ithlete smart phone application for determining ultra-short-term heart rate variability. Journal of Human Kinetics, 39(1):85–92, 2013.

[51] D. French and L. Torres Ronda, editors. NSCA's Essentials of Sport Science. Human Kinetics, 1 edition, 2022.

[52] J Friel. Total Heart Rate Training: Customize and Maximize Your Workout Using a Heart Rate Monitor. Berkeley: Ulysses Press, 13 edition, 2006.

[53] T.-C. Fu. A review on time series data mining. Engineering Applications of Artificial Intelligence, 24:164–181, 2011.

[54] D. Fuller, E. Colwell, J. Low, K. Orychock, M.A. Tobin, B. Simango, R. Buote, D. Van Heerden, H. Luan, Ki. Cullen, L. Slade, and N.G.A. Taylor. Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: Systematic review. JMIR mHealth and uHealth, 8(9):e18694, 2020.

[55] J.C.B. Gamboa. Deep learning for time-series analysis. arXiv, (1701.01887), 2017.

[56] K.E. Georgiou, R.K. Dimov, N.B. Boyanov, K.G. Zografos, A.V. Larentzakis, and B.I. Marinov. Feasibility of a new wearable device to estimate acute stress in novices during high-fidelity surgical simulation. Folia Med, 61(49–60), 2019.

[57] E Gil, M Orini, R Bailón, J M Vergara, L Mainardi, and P Laguna. Photoplethysmography pulse rate variability as a surrogate measurement of heart rate variability during non-stationary conditions. Physiological Measurement, 31(9):1271–1290, 2017.

[58] S. Gillinov, M. Etiwy, R. Wang, G. Blackburn, D. Phelan, A. M. Gillinov, P. Houghtaling, H. Javadikasgari, and M.Y. Desai. Variable accuracy of wearable heart rate monitors during aerobic exercise. Medicine and Science in Sports and Exercise, 49(8):1697–1703, 2017.

[59] P. Gupta, D.a Agrawal, J. Chhabra, and P.K. Dhir. iot based smart healthcare kit. International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), 2016:237–242, 2016.

[60] S. Gupta and L. Wang. Stock forecasting with feedforward neural networks and gradual data sub-sampling. Australian Journal of Intelligent Information Processing Systems, 11(4), 2010.

[61] E. C. Hart, N. Charkoudian, and V. M. Miller. Sex, hormones and neuroeffector mechanisms: Sex, hormones and neurotransmission. Acta Physiologica, 203(1):155–165, 2011.

[62] P. Haynie. Guidance for industry. `https://www.fda.gov/downloads/drugs/guidances/ucm070287.pdf`.

[63] E. Hermand, J. Cassirame, G. Ennequin, and O. Hue. Validation of a photoplethysmographic heart rate monitor: Polar oh1. Int J Sports Med, 40:462–7, 2019.

[64] D. Hernando, N Garatachea, R. Almeida, J.A. Casajus, and R. Bailon. Validation of heart rate monitor polar rs800 for heart rate variability analysis during exercise. Journal of Strength and Conditioning Research, 32(3):716–725, 2018.

[65] I.T. Hettiarachchi, S. Hanoun, and D. Nahavandi. Validation of polar oh1 optical heart rate sensor for moderate and high intensity physical activities. PLoS One, 14:e0217288, 2019.

[66] J.E. Horton, P. Stergiou, T.S. Fung, and L. Katz. Comparison of polar m600 optical heart rate and ecg heart rate during exercise. Medicine and Science in Sports and Exercise, 49:2600–7, 2017.

[67] T. Ince, S. Kiranyaz, L. Eren, M. Askar, and M. Gabbouj. Real-time motor fault detection by 1-d convolutional neural networks. IEEE Transactions on Industrial Electronics, 63(11):7067–7075, 2012.

[68] E. Jo, K. Lewis, D. Directo, M.J. Kim, and B.A. Dolezal. Validation of biofeedback wearables for photoplethysmographic heart rate tracking. Journal of Sports Science Medicine, 15(3):540–547, 2016.

[69] A.D. Jose and D. Collison. The normal range and determinants of the intrinsic heart rate in man. Cardiovascular Research, 4(2):160–167, 1970.

[70] N.K. Kakria, P.and Tripathi and P. Kitipawang. A real-time health monitoring system for remote cardiac patients using smartphone and wearable sensors. International Journal of Telemedicine and Applications, 8:1–11, 2015.

[71] A. Kamal, J. Harness, G. Irving, and A. Mearns. Skin photoplethysmography review. Computer methods and programs in biomedicine, 28(4):257–269, 1989.

[72] A. Kashou, A. May, C. DeSimone, and P. Noseworthy. The essential skill of ecg interpretation: How do we define and improve competency? Postgraduate Medical Journal, 96(1133):125–127, 2020.

[73] A. Katrompas, T. Ntakouris, and V. Metsis. Recurrence and self-attention vs the transformer for time-series classification: A comparative study. In Martin Michalowski, Syed Sibte Raza Abidi, and Samina Abidi, editors, Artificial Intelligence in Medicine, volume 13263, pages 99–109. Springer International Publishing.

[74] L.R. Keytel, J.H. Goedecke, T.D. Noakes, H. Hiiloskorpi, R. Laukkanen, L. Van Der Merwe, and E.V. Lambert. Prediction of energy expenditure from heart rate monitoring during submaximal exercise. Journal of Sports Sciences, 23(3):289–297, 2005.

[75] M. Khan, A. Khan, AS. Al-Mubarak, and MS. Khan. Clinically insignificant differences in heart rate derived from three devices: a prospective observational study. Ann Intern Med, 169(11):825–832, 2018.

[76] M. Khashei, M. Bijari, and S.R. Hejazi. Combining seasonal arima models with computational intelligence techniques for time series forecasting. Soft Computing, 16:1091–1105, 2012.

[77] S.-H. Kim and E.-R. Jeong. 1-dimentional convolutional neural network based heart rate estimation using photoplethysmogram signals. Webology, 19(1):4571–4580, 2022.

[78] Y. Kim, C.. Denton, L. Hoang, and A.M. Rush. Structured attention networks. arXiv, (1702.00887), 2017.

[79] S. Kiranyaz, T. Ince, and M. Gabbouj. Real-time patient-specific ecg classification by 1-d convolutional neural networks. IEEE Transactions on Biomedical Engineering, 63(3):664–675, 2016.

[80] Ensio-Lehtonen A. Kitchenham, S. Human-data interaction in healthcare. In Human-computer interaction in healthcare. IGI Global, 2020.

[81] R.R. Kroll, J.G. Boyd, and D.M. Maslove. Accuracy of a wrist-worn wearable device for monitoring heart rates in hospital inpatients: A prospective observational study. Journal of Medical Internet Research, 18:e253, 2016.

[82] R. Kruse, C. Borgelt, C. Braune, S. Mostaghim, and M. Steinbrecher. Computational Intelligence. Springer London.

[83] L Landsberg and J. Young. Effects of nutritional status on autonomic nervous system function. J Clin Nutr, 35:1234–1240, 1982.

[84] J Lee and R.G. Mark. A hypotensive episode predictor for intensive care based on heart rate and blood pressure time series. Computing in Cardiology, page 8184, 2010.

[85] C. Li and J.W. Hu. A new arima-based neuro-fuzzy approach and swarm intelligence for time series forecasting. Journal of Engineering Applications of Artificial Intelligence, 25:295–308, 2012.

[86] H. Li. Asynchronism-based principal component analysis for time series data mining. Expert Systems with Applications, 41(6):2842–2850, 5 2014.

[87] H. Li. On-line and dynamic time warping for time series data mining. International Journal of Machine Learning and Cybernetics, 6(1):145–153, 2015.

[88] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, and X. Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. (1907.00235), 2020.

[89] X. Li, S. Wu, and L. Wang. Blood pressure prediction via recurrent models with contextual layer. Proceedings of the 26th International Conference on World Wide Web, pages 685–693, 2017.

[90] M. Liu, S. Ren, S. Ma, J. Jiao, Y. Chen, Z. Wang, and W. Song. Gated transformer networks for multivariate time series classification. arXiv, (2103.14438), 2021.

[91] I.E. Livieris, E. Pintelas, and P. Pintelas. A cnn-lstm model for gold price time-series forecasting. Neural Computing and Applications, 32(23):17351–17360, 2020.

[92] I. López-Yáñez, L. Sheremetov, and C. Yáñez-Márquez. A novel associative model for time series data mining. Pattern Recognition Letters, 41:23–33, 5 2014.

[93] G. Lu, F. Yang, J. A. Taylor, and J. F. Stein. A comparison of photoplethysmography and ecg recording to analyse heart rate variability in healthy subjects. Journal of Medical Engineering Technology, 33(8):634–641, 2009.

[94] Y. Maeda, M. Sekine, and T. Tamura. The advantages of wearable green reflected photoplethysmography. Journal of Medical Systems, 35(5):829–834, 2011.

[95] Y. Maeda, M. Sekine, and T. Tamura. Relationship between measurement site and motion artifacts in wearable reflected photoplethysmography. Journal of Medical Systems, 35(5):969–976, 2011.

[96] P. Mandal, T. Senjyu, N. Urasaki, A. Yona, Funabashi T., and Srivastava A.K. Price forecasting for day-ahead electricity market using recursive neural network. IEEE Power Engineering Society General Meeting, pages 1–8, 2007.

[97] S. Masum, J.P. Chiverton, Y. Liu, and B. Vuksanovic. Investigation of machine learning techniques in forecasting of blood pressure time series data. International Conference on Innovative Techniques and Applications of Artificial Intelligence, 11927:269–282, 2019.

[98] Priv.-Doz. Dr. Dr. med. Mahdi Sareban, Salzburger Landeskliniken, Ludwig Boltzmann Institute for Digital Health, and Prevention. `https://www.clinicaltrials.gov/study/NCT05525000?locStr=Salzburg,%20Austria&country=Austria&state=Salzburg&city=Salzburg&distance=50&term=ValOpti&rank=1`, 2023.

[99] L. Menghini, E. Gianfranchi, N. Cellini, E. Patron, M. Tagliabue, and M. Sarlo. Stressing the accuracy: Wristworn wearable sensor validation over different conditions. Psychophysiology, 56:e13441, 2019.

[100] S. Michael, O. Jay, M. Halaki, K. Graham, and GM Davis. Submaximal exercise intensity modulates acute post-exercise heart rate variability. Eur J Appl Physiol, 116:697–706, 2016.

[101] S.M. Molaei and M.R. Keyvanpour. An analytical review for event prediction system on time series. 2nd International Conference on Pattern Recognition and Image Analysis, pages 1–6, 2015.

[102] J.M. Mühlen, J. Stang, Lykke S.E., P.B. Judice, P. Molina-Garcia, W. Johnston, L.B. Sardinha, F.B. Ortega, B. Caulfield, W. Bloch, S. Cheng, U. Ekelund, J.C. Brønd, A. Grøntved, and M. Schumann. Recommendations for determining the validity of consumer wearable heart rate devices: expert statement and checklist of the INTERLIVE network. British Journal of Sports Medicine, 55(14):767–779, 2021.

[103] A.M. Müller, N.X. Wang, J. Yao, C. Seng Tan, I.C.C. Low, N. Lim, J. Tan, A. Tan, and F. Müller-Riemenschneider. Heart rate measures from wrist-worn activity trackers in a laboratory and free-living setting: Validation study. JMIR mHealth and uHealth, 7:e14120, 2019.

[104] B. Murugesan, V. Ravichandran, K. Ram, S.P. Preejith, J. Jayaraj, S.M. Shankaranarayana, and M. Sivaprakasam. Ecgnet: Deep network for arrhythmia classification. IEEE International Symposium on Medical Measurements and Applications (MeMeA), pages 1–6, 2018.

[105] J.W. Navalta, J. Montes, N.G. Bodell, R.W. Salatto, J.W. Manning, and M. DeBeliso. Concurrent heart rate validity of wearable technology devices during trail running. PLoS One, 15(8):e0238569, 2020.

[106] B.W. Nelson and N.B. Allen. Accuracy of consumer wearable heart rate measurement during an ecologically valid 24-hour period: Intraindividual validation study. JMIR mHealth and uHealth, 7(3):e10828, 2019.

[107] M. Nielsen. Neural networks and deep learning. Determination Press, 2018.

[108] L. Niu, C. Chen, H. Liu, S. Zhou, and M. Shu. A deep-learning approach to ecg classification based on adversarial domain adaptation. Healthcare, 8(4):437, 2020.

[109] P.A. Obrist, Webb R.A., J.R. Sutterer, and J.L. Howard. The cardiac-somatic relationship: some reformulations. Psychophysiology, 6:569–587, 1970.

[110] F. Olaiya and A.B. Adeyemo. Application of data mining techniques in weather prediction and climate change studies. International Journal of Information Engineering and Electronic Business, 4(1):51–59, 2012.

[111] F. Ongenae, S. Van Looy, D. Verstraeten, D. Verplancke, T.and Benoit, F. De Turck, T. Dhaene, B. Schrauwen, and J. Decruyenaere. Time series classification for the prediction of dialysis in critically ill patients using echo statenetworks. Engineering Applications of Artificial Intelligence, 26(3):984–996, 2013.

[112] M. Oyeleye, T. Chen, S. Titarenko, and G. Antoniou. A predictive analysis of heart rates using machine learning techniques. Int J Environ Res Public Health, 19(4):2417, 2022.

[113] J. Parak and I. Korhonen. Evaluation of wearable consumer heart rate monitors based on photopletysmography. 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pages 3670–3673, 2014.

[114] J. Parak, M. Uuskoski, J. Machek, and I. Korhonen. Estimating heart rate, energy expenditure, and physical performance with a wrist photoplethysmographic device during running. JMIR mHealth and uHealth, 5(7):e97, 2017.

[115] S. Passler, N. Müller, and V. Senner. In-ear pulse rate measurement: A valid alternative to heart rate derived from electrocardiography? Sensors, 19(3641):8, 2019.

[116] M.S. Patel, D. Asch, and K. Volpp. Wearable devices as facilitators, not drivers, of health behavior change. The Journal of the American Medical Association, 313(5):459–60, 2014.

[117] G. Pelizzo, A. Guddo, A. Puglisi, A. De Silvestri, C. Comparato, M. Valenza, E. Bordonaro, and V. Calcaterra. Accuracy of a wrist-worn heart rate sensing device during elective pediatric surgical procedures. Children, 5(38), 2018.

[118] V. Pichot, F. Roche, J.-M. Gaspoz, F. Enjolras, A. Antoniadis, P. Minini, F.c Costes, T. Busso, J.-R. Lacour, and J.C. Barthelemy. Relation between heart

rate variability and training load in middle-distance runners:. Medicine and Science in Sports and Exercise, 32(10):1729–1736, 2000.

[119] MF Piepoli, AW Hoes, S Agewall, C Albus, C Brotons, AL Catapano, MT Cooney, U Corrà, B Cosyns, C Deaton, I Graham, MS Hall, FDR Hobbs, ML Lochen, H Löllgen, P Marques-Vidal, J Perk, E Prescott, J Redon, DJ Richter, N Sattar, Y Smulders, M Tiberi, HB van der Worp, I van Dis, WMM Verschuren, and S Binno. 2016 european guidelines on cardiovascular disease prevention in clinical practice: The sixth joint task force of the european society of cardiology and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of 10 societies and by invited experts)developed with the special contribution of the european association for cardiovascular prevention rehabilitation (eacpr). Eur Heart J., 37(29):2315–2381, 2016.

[120] D.J. Plews, P.B. Laursen, A.E. Kilding, and M. Buchheit. Heart rate variability in elite triathletes, is variation in variability the key to effective training? a case comparison. European Journal of Applied Physiology, 112(11):3729–3741, 2012.

[121] Y. Qiu, Y. Liu, J. Arteaga-Falconi, H. Dong, and A.E. Saddik. Evm-cnn: Real-time contactless heart rate estimation from facial video. IEEE Transactions on Multimedia, 21(7):1778–1787, 2019.

[122] N.F.M. Radzuan, Z. Othman, and A.A. Bakar. Uncertain time series in weather prediction. Procedia Technology, 11:557–564, 2013.

[123] S. Raschka and V. Mirjalili. Python machine learning: machine learning and deep learning with Python, scikit-learn, and TensorFlow 2. Expert insight. Packt.

[124] A. Ravizza, L. Di Pietro, F. Sternini, C. De Maria, C. Bignardi, and A. Audenino. Comprehensive review on current and future regulatory requirements on wearable sensors in preclinical and clinical testing. Front Bioeng Biotechnol, 7:313, 2019.

[125] R.K. Reddy, R. Pooni, D.P. Zaharieva, B. Senf, J. El Youssef, E. Dassau, Francis J. Doyle I., M.A. Clements, M.R. Rickels, S.R. Patton, J.R. Castle, M.C. Riddell, and P.G. Jacobs. Accuracy of wrist-worn activity monitors during common daily physical activities and types of structured exercise: Evaluation study. JMIR mHealth and uHealth, 6(12):e10338, 2018.

[126] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven. Deep ppg: Large-scale heart rate estimation with convolutional neural networks. Sensors,

19(14):3079, 2019.

[127] A.C. Rencher and G.B. Schaalje. Linear models in statistics. Wiley-Interscience, 2nd ed edition. OCLC: ocn144331522.

[128] S. Rendle. Factorization machines. IEEE International Conference on Data Mining, pages 995–1000, 2010.

[129] M. Rhudy. Time alignment techniques for experimental sensor data. IJCSES, 5:1–14, 2014.

[130] B.F. Robinson, S.E. Epstein, G.D. Beiser, and E. Braunwald. Control of heart rate by the autonomic nervous system: Studies in man on the interrelation between baroreceptor mechanisms and exercise. Circulation Research, 19(2):400–411, 1966.

[131] J.G. Rousselle, J. Blascovich, and R.M. Kelsey. Cardiorespiratory response under combined psychological and exercise stress. Psychophysiol, 20:49–58, 1995.

[132] F. Sartor, G. Papini, Lieke G.E. Cox, and J. Cleland. Methodological shortcomings of wrist-worn heart rate monitors validations. Journal of Medical Internet Research, 20(7):e10108, 2018.

[133] T. Schack, M. Muma, and A.M. Zoubir. Computationally efficient heart rate estimation during physical exercise using photoplethysmographic signals. 25th European Signal Processing Conference, pages 2478–2481, 2017.

[134] G.A.F. Seber and A.J. Lee. Linear Regression Analysis. A John Wiley Sons Publication, 2 edition, 2003.

[135] H. Selye. The evolution of the stress concept: The originator of the concept traces its development from the discovery in 1936 of the alarm reaction to modern therapeutic applications of syntoxic and catatoxic hormones. American Scientist, 61(6):692–699, 1973.

[136] A. Shcherbina, C. Mattsson, D. Waggott, H. Salisbury, J. Christle, T. Hastie, M. Wheeler, and E. Ashley. Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. Journal of Personalized Medicine, 7(2):3, 2017.

[137] M. Shen, L. Zhang, X. Luo, and J. Xu. Atrial fibrillation prediction algorithm based on attention model. Journal of Physics: Conference Series, 1575(1):012122, 2020.

[138] R.H. Shumway and D.S. Stoffer. Time Series Analysis and Its Applications with R examples. Springer New York, 2006.

[139] A. Shyam, V. Ravichandran, S.P. Preejith, Jayaraj J., and S. Mohanasankar. Ppgnet: Deep network for device independent heart rate estimation from photoplethysmogram. IEEE Engineering in Medicine and Biology Society Conference, (1903.08912), 2019.

[140] L.E.V. Silva, H.T. Moreira, M.M.M. Bernardo, A. Schmidt, M.M.D. Romano, H.C.r Salgado, R. Fazan, R. Tinós, and A. Marin-Neto. Prediction of echocardiographic parameters in chagas disease using heart rate variability and machine learning. Biomedical Signal Processing and Control, 67, 5 2021.

[141] A.J. Smola and B. Schölkopf. A tutorial on support vector regression. Statistics and Computing, 14(3):199–222, 2004.

[142] K.E. Speer, N. Semple, S. Naumovski, and A.J. McKune. Measuring heart rate variability using commercially available devices in healthy children: A validity and reliability study. Eur. J. Investig. Health Psychol. Educ., 10(1):390–404, 2020.

[143] R. Špetlík, V. Franc, J. Cech, and Matas. J. Visual heart rate estimation with convolutional neural network. BMVC, 2018.

[144] S. Sreejith, S. Rahul, and R. C. Jisha. A real-time patient monitoring system for heart disease prediction using random forest algorithm. Advances in Signal Processing and Intelligent Recognition Systems, pages 485–500, 2015.

[145] S.E. Stahl, H.-S. An, D.M. Dinkel, J.M. Noble, and J.-M. Lee. How accurate are the wrist-based heart rate monitors during walking and running activities? are they accurate enough? BMJ Open Sport Exerc Med, 2(1):e000106, 2016.

[146] J. Stanley, J.M. Peake, and M. Buchheit. Cardiac parasympathetic reactivation following exercise: Implications for training prescription. Sports Medicine, 43(12), 2013.

[147] M.P. Stove, E. Haucke, M.L. Nymann, T. Sigurdsson, and B.T. Larsen. Accuracy of the wearable activity tracker garmin forerunne r235 for the assessment of heart rate during rest and activity. JSports Sci, 37(8):895–901, 2019.

[148] Y. Sun and N. Thakor. Photoplethysmography revisited: From contact to non-contact, from point to imaging. IEEE Transactions on Biomedical Engineering, 63(3):463–477.

[149] G. Swaptna, K.P. Soman, and V. Ravi. Automated detection and classifying diabetes mellitus using cnn and cnn-lstm network and heart rate signals. Procedia Computer Science, 132:1253–1262, 2018.

[150] Fu T.-C. A review on time series data mining. Engineering Applications of Artificial Intelligence, 24(1):164–181, 2 2011.

[151] T. Tamura, Y. Maeda, M. Sekine, and M. Yoshida. Wearable photoplethysmographic sensors - past and present. Electronics, 3(2):282–302, 2014.

[152] G.T. Taye, H.-J. Hwang, and K.M. Lim. Application of a convolutional neural network for predicting the occurrence of ventricular tachyarrhythmia using heart rate variability features. Scientific Reports, 10(1):6769, 2020.

[153] R.S. Thiebaud, M.D. Funk, J.C. Patton, B.L. Massey, T.E. Shay, M.G. Schmidt, and N. Giovannitti. Validity of wrist-worn consumer products to measure heart rate and energy expenditure. Digital Health, 4, 2018.

[154] W.R. Thompson. Worldwide survey of fitness trends for 2022. ACSMs Health Fitness Journal, 26(1):11–20, 2022.

[155] G.J. Tortora and B.H. Derrickson. Principles of Anatomy and Physiology, volume 1. John Wiley Sons, 12 edition, 2009.

[156] Y. Tran, N. Wijesuriya, M. Tarvaining, P. Karjalainen, and A. Craig. The relationship between spectral changes in heart rate variability and fatigue. Psychophysiol, 23:143–151, 2009.

[157] S. Tuli, G. Casale, and N.R. Jennings. Tranad: Deep transformer networks for anomaly detection in multivariate time series data. arxiv, (2201.07284), 2022.

[158] S. Walczak and N. Cerpa. Artificial neural networks. Encyclopedia of Physical Science and Technology, pages 631–645, 2003.

[159] L. K. Wallace, K. M. Slattery, and Aaron J. Coutts. A comparison of methods for quantifying training load: relationships between modelled and actual training responses. European Journal of Applied Physiology, 114(1):11–20, 2014.

[160] M.P. Wallen, S.R. Gomersall, S.E. Keating, U. Wisløff, and J.S. Coombes. Accuracy of heart rate watches: Implications for weight management. PLoS One, 11(5):e0154420, 2016.

[161] R. Wang, G. Blackburn, M. Desai, D. Phelant, L. Gillinov, P. Houghtaling, and M. Gillinov. Accuracy of wrist-worn heart rate monitors. JAMA Cardiology, 2(1):104, 2017.

[162] D.E.R. Warburton, C.W. Nicol, and S.S. Bredin. Health benefits of physical activity: the evidence. Canadian Medical Association Journal, 174(6):801–809, 2006.

[163] P. Warrick and E. Hamilton. Lstm modeling of perinatal fetal heart rate. Computing in Cardiology, 46:1–4, 2019.

[164] M Weippert, M. Kumar, S. Kreuzfeld, D. Arndt, A. Rieger, and R. Stoll. Comparison of three mobile devices for measuring r-r intervals and heart rate variability: Polar s810i, suunto t6 and an ambulatory ecg system. Eur J Appl Physiol, 109(4):779–86, 2010.

[165] Q. Wen, T. Zhou, C. Zhang, Z. Chen, W.and Ma, J. Yan, and L. Sun. Transformers in time series: A survey. arXiv, (2202.07125), 2022.

[166] S. Williams, T. Booton, M. Watson, D. Rowland, and M. Altini. Heart rate variability is a moderating factor in the workload-injury relationship of competitive CrossFitTM athletes. Journal of Sports Science Medicine, 16:443–449, 2017.

[167] K. Wilson, C. Bell, L. Wilson, and H. Witteman. Agile research to complement agile development: a proposal for an mHealth research lifecycle. NPJ Digital Medicine, 1(1):46, 2018.

[168] C. Woo, G.vand Liu, D. Sahoo, A. Kumar, and S. Hoi. Deeptime: Deep time-index meta-learning for non-stationary time-series forecasting. arXiv, (2207.06046).

[169] Z. Xiao, X. Xu, H. Zhang, and E. Szczerbicki. A new multi-process collaborative architecture for time series classification. Knowledge-Based Systems, 220:106934, 2021.

[170] D. Xie, J.vand Wen, L. Liang, L. Jia, Y.and Gao, and J. Lei. Evaluating the validity of current mainstream wearable devices in fitness tracking under various physical activities: Comparative study. JMIR mHealth and uHealth, 6(4):e94, 2018.

[171] J. Xu, H. Wu, J. Wang, and M. Long. Anomaly transformer: Time series anomaly detection with association discrepancy. arXiv, (2110.02642), 2022.

[172] Y.-Y. Yang, C.-H. H.and Tsai and P.-Y. Chen. Voice2series: Reprogramming acoustic models for time series classification. arXiv, (2106.09296), 2022.

[173] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff. A transformer-based framework for multivariate time series representation learning. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pages 2114–2124, 2021.

[174] H. Zhang. An improved qrs wave group detection algorithm and matlab implementation. Physics Procedia, 25:1010–1016, 2012.

[175] Y. Zhang, Y. Ning, Z. Huan, B. Li, and Y. Liu. Short-term heart rate prediction approach based on cnn-gru model with an attention mechanism. Journal of Physics: Conference Series, 2030(1):012060, 2021.

[176] Z. Zhang, Z. Pi, and B. Liu. Troika: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise. IEEE Transactions on Biomedical Engineering, 62(2):522–531, 2015.

[177] B. Zhao, H. Lu, S. Chen, J. Liu, and D. Wu. Convolutional neural networks for time series classification. Journal of Systems Engineering and Electronics, 28(1):162 – 169, 2 2017.

[178] H. Zhao, J. Jia, and V. Koltun. Exploring self-attention for image recognition. arXiv, (2004.13621), 2020.

[179] M. Zhu and L. Wang. Intelligent trading using support vector regression and multilayer perceptrons optimized with genetic algorithms. International Joint Conference on Neural Networks, pages 1–5, 2010.

Appendix

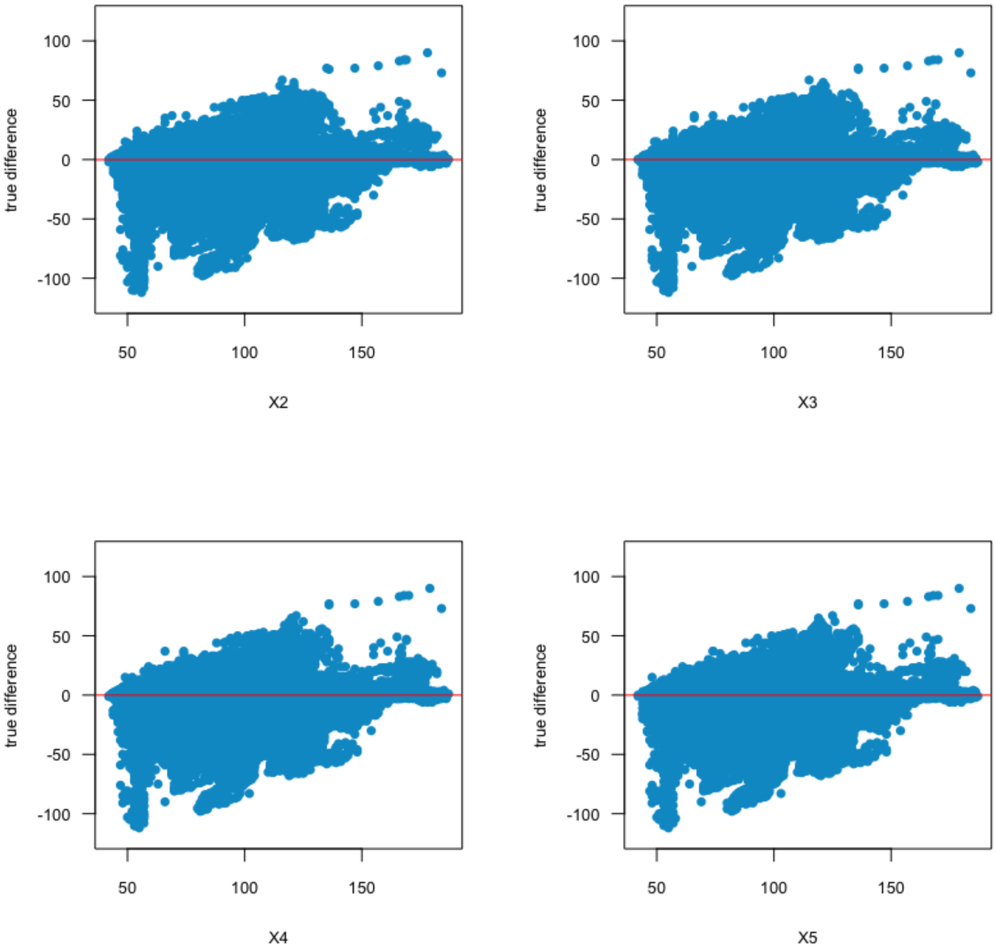


Figure 7.1: The remaining variables of the Multiple Linear Regression (MLR) model I.

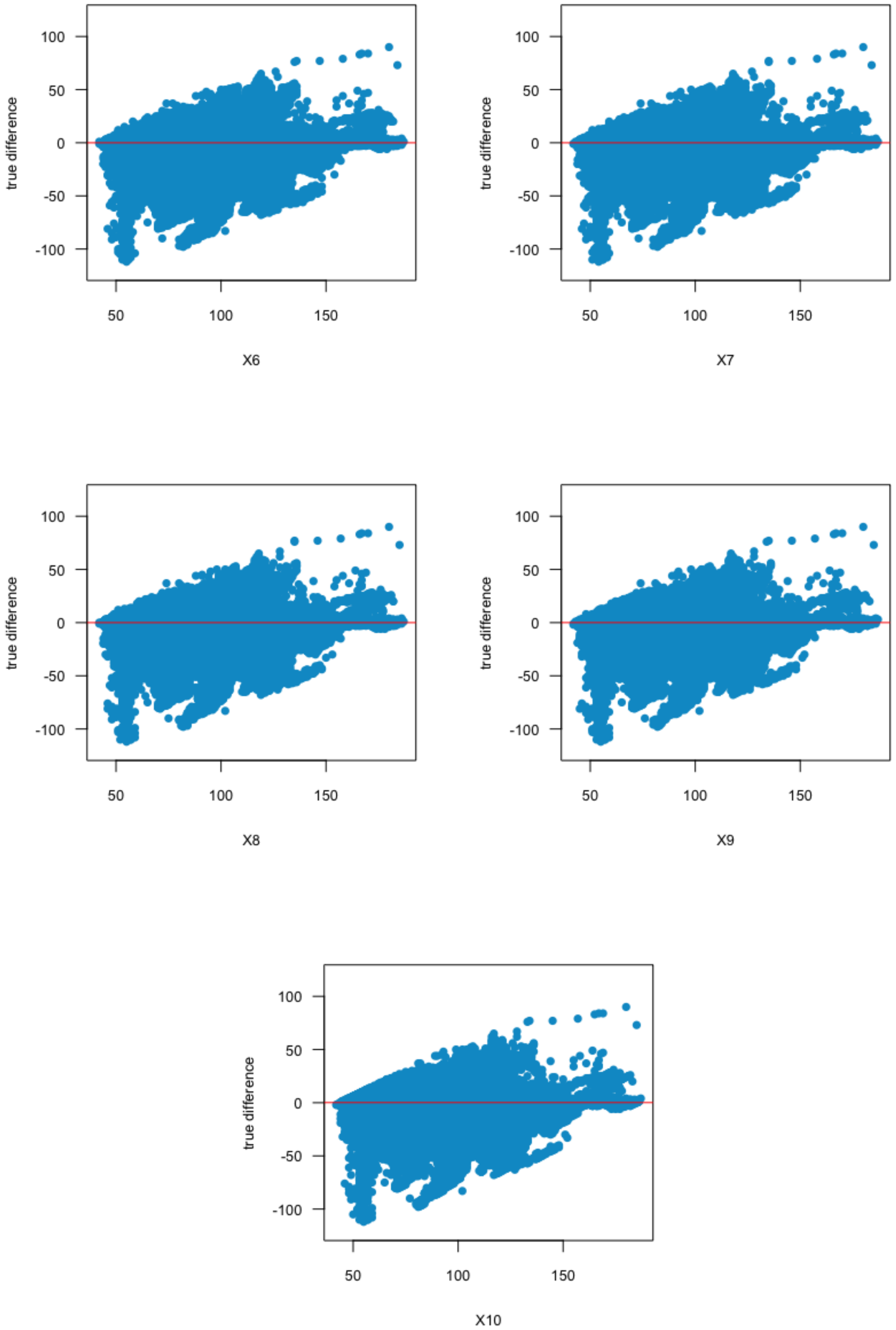


Figure 7.2: The remaining variables of the Multiple Linear Regression (MLR) model II.