Machine Learning
Data Analytics

FAU Friedrich-Alexander-Universität
Technische Fakultät

Improving the Robustness of Radar-Based Heart Sound Detection

Bachelor's Thesis in Medical Engineering

submitted
by

Clark Bäker

born 28.12.1999 in Hamburg

Written at

Machine Learning and Data Analytics Lab
Department Artificial Intelligence in Biomedical Engineering
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

in Cooperation with

TU Hamburg

Advisors: Luca Abel M. Sc., Nils Albrecht M. Sc., Daniel Krauß M. Sc., Robert Richer M. Sc., Prof. Dr. Bjoern Eskofier, Prof. Dr. Alexander Kölpin, Prof. Dr. Nicolas Rohleder

Started: 02.05.2023

Finished: 02.10.2023

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.
Die Richtlinien des Lehrstuhls für Bachelor- und Masterarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Erlangen, den 28. September 2023

Übersicht

Die berührungslose Überwachung der Vitalparameter ermöglicht eine kontinuierliche und bequeme Datenerfassung, die die Unannehmlichkeiten der herkömmlichen kontaktbasierten Messung verringert. Die radarbasierte Erkennung von Herztönen ist ein aufstrebender Bereich bei der nicht-invasiven Messung der Vitalparameter einer Person. Die Segmentierung der Herztöne bietet Möglichkeiten zur kontaktlosen Beurteilung wichtiger Vitalparameter wie der Herzfrequenz oder der Herzfrequenzvariabilität. Eine hochmoderne Methode zur Erkennung grundlegender Herztöne aus Radaraufzeichnungen verwendet ein neuronales Netzwerk, das auf bidirektionalen Kurzzeitgedächtniszellen (biLSTM) basiert. Obwohl sich diese Architektur in Umgebungen mit wenig stattfindender Bewegung bewährt hat, wurde sie noch nicht systematisch evaluiert, wenn sie auf Daten mit zufälligen großen Körperbewegungen trainiert und getestet wurde.

In dieser Arbeit wird ein Algorithmus zur modernen Detektion von Herzgeräuschen vorgestellt, der in ein Open-Source-Framework implementiert wurde, das am Lehrstuhl für Maschinelles Lernen und Datenanalytik der Friedrich-Alexander-Universität Erlangen-Nürnberg entwickelt wird. Die Beurteilung des Algorithmus erfolgt mit Daten, welche zufällige Bewegungen von sitzenden Teilnehmern enthalten, die einen Stresstest mit variierender Bewegungsintensität durchführten. Insbesondere wurde ein neuronales biLSTM-Netzwerk trainiert, um den ersten Herzton im Radar zu erkennen. Die Radarmessungen sind mit einer Front- und einer Rücken-fokussierten Antenne aufgezeichnet worden. In einem ersten Schritt wurde die Modellkonfiguration im Hinblick auf die Komplexität des neuronalen Netzes und die Höhe des Dropouts zwischen den verschiedenen Schichten optimiert. Anschließend wurde die Leistung des resultierenden Modells über verschiedene Versuchsphasen hinweg und zwischen den verwendeten Radarsensoren verglichen. Die erzielten Ergebnisse zeigen, dass der Algorithmus über alle getesteten Varianten hinweg schwache Leistung mit nur geringen Schwankungen aufweist. Für eine Toleranz von ± 100 ms um den Herzton wurde mit den Daten der Dorsalantenne über alle Versuchsphasen ein maximaler F1-Score von 0.48 ± 0.05 erreicht. Die Leistung konnte durch die Verwendung der Frontalantenne auf einen F1-Score von 0.60 ± 0.19 verbessert werden. Außerdem war die Leistung in Studienphasen geringerer Bewegung deutlich höher und erreichte einen F1-Wert von 0.74 ± 0.22.

Eine wichtige Erkenntnis dieser Arbeit ist, dass die ausschließliche Verwendung von Trainingsdaten, die Bewegungsartefakte enthalten, keine Resistenz gegen solche Artefakte gewährleistet. Eine wesentliche Verbesserung der Genauigkeit bei der Erkennung von Herztönen konnte durch die Verwendung einer Radarantenne erreicht werden, die auf die Vorderseite statt auf die Rückseite einer sitzenden Person gerichtet ist. Eine größere Robustheit gegenüber Bewegungen könnte jedoch nur durch die Erfassung und Verarbeitung besserer Daten erreicht werden.

Abstract

Contactless monitoring of vital signs provides continuous and convenient data collection reducing inconveniences of traditional contact-based sensing. Radar-based detection of heart sounds is an emerging field in the non-invasive measurement of an individual's vital signs. The segmentation of heart sounds offers possibilities for the remote assessment of important vital parameters such as the heart rate (HR) or the heart rate variability (HRV). One state-of-the-art method to detect the event of fundamental heart sounds from radar recordings utilizes a neural network based on bi-directional long short-term memory (biLSTM) cells. Although proving itself in settings with little movement by the person under investigation, this architecture has not yet been systematically evaluated when trained and tested on data containing random large body movements.

To fill this gap, this work introduces a state-of-the art heart sound detection pipeline included in an open-source framework developed at the Machine Learning and Data Analytics lab at the Friedrich-Alexander Universität Erlangen-Nürnberg and evaluated its performance on data containing random motion from sitting participants performing a stress test. In particular, a biLSTM neural network was trained to detect the first fundamental heart sound using radar recorded by one dorsal and one frontal sensor during the study encompassing phases of more or less movement. As a first step the model configuration was optimized with respect to the complexity of the neural network and the amount of dropout between the different layers. Afterward the performance of the final model was compared across different experimental phases and among different radar sensors used.

The results obtained in this work showed that the developed pipeline is performing weakly across all tested hyperparameter configurations with only slight variation. For a tolerance window of ± 100 ms around the ground truth heart sound a maximum F1-score of 0.48 ± 0.05 was achieved over all experimental phases using the data of the dorsal antenna. However, performance could be considerably improved by using the frontal antenna to a F1-score of 0.60 ± 0.19. Furthermore, the performance was much higher in phases of less movement, reaching up to a F1-score of 0.74 ± 0.22.

A key finding of this work is that depending solely on data containing motion artifacts for model training does not ensure its resistance to such artifacts. Substantial improvement regarding the heart sound detection accuracy was achieved using a radar antenna focusing a sitting person's front instead of back. However, greater robustness might only be possible with the acquisition and processing of superior data.

Contents

Chapter 1

Introduction

Continuous monitoring of vital signs, including heart rate (HR), respiration rate, blood pressure and others, is crucial for monitoring and predicting the physiological status of individuals [Bre19]. Contactless vital sign monitoring is desirable in various healthcare settings to avoid the inconvenience and discomfort of traditional contact-based sensing [Wen21]. Further in hazardous environments like battlefields or mountain accidents it may be critical to know if a person shows vital signs or not, without the need to be in direct contact [Ord18].

To this end, radar technology emerges as a promising and transformative way to assess cardiac activity. Continuous wave Doppler radar (CWDR) can measure the relative displacement of the human body resulting from the contraction of the heart muscle. Additionally, radiofrequency waves have the advantage over, e.g. laser, to penetrate various materials such as clothing, bedding or snow making the remote sensing widely applicable in healthcare and other domains.

Previous research in this field primarily aimed to retrieve heartbeats from the low-frequency components present in the radar measurement. These components correspond to the pulse wave emerging from the muscular action of the heart [Li13; Xio17; Tu16]. A novel approach to analyze cardiac activity is centered around the segmentation of the fundamental heart sounds (FHSs) from the radar. The origin of the FHS can be traced to the closure of the heart valves during each cardiac cycle, with the first sound occurring when the atrioventricular valves close at the beginning of systole and the second sound resulting from the closure of the semilunar valves at the end of systole. Will et al. have pioneered the field of radar-based heart sound detection by linking vibration on the body's surface caused by these closures of the heart valves to higher frequency components in the measured radar [Wil18]. They presented a hidden semi-Markov model (HSMM) that could reliably segment the heart sounds within the radar data and following this lead, Shi and colleagues achieved the same with a bidirectional long short-term memory (biLSTM) network [Shi19]. Further, the determination of very precise interbeat intervals (IBIs) based on the segmented heart sounds was proven to monitor a person's heart rate variability (HRV) with a relative error of only around 5% offering great possibilities to observe autonomic nervous system (ANS) activity remotely in the future [Shi21].

Thus far, these accomplishments have been limited to radar recordings in settings with minimal to no movement from the person being tested. A recent study examined the effect of sensor positioning and semi-standardized movements, which resulted in considerable inaccuracies of the HSMM trained on motionless data across all sensor positions [Her22]. Additionally, the datasets that have been published so far to support developments in the field, containing radar and reference signals such as electrocardiogram (ECG) or phonocardiogram (PCG), do not include movements that can be realistically expected, i.e. random large body movements (RLBM).

Therefore, this thesis aims to acquire and curate a dataset comprising realistic data for testing motion-robust analysis pipelines in the field. To accomplish this, the study collected radar data and reference ECG from 45 participants who underwent both the Trier Social Stress Test (TSST) for acute psychosocial stress induction [Kir93] and a modified control version, the f-TSST [Wie13], on two consecutive days. The raw radar data was recorded using either 2 or 4 antennae, depending on whether the participant was seated or standing, respectively. Afterwards, the hyperparameters of a heart-sound segmentation pipeline using a biLSTM will undergo a grid search. Thereby the tested parameter configurations' influence on the pipeline's resistance against motion artifacts will be investigated in this work. Further, the training and validation of the different pipelines rely solely on data from the study, thus incorporating RLBM. This works aims to extensively analyze the influence of incorporating data from realistic scenarios into training and validation of the used approach. Finally, the evaluation will encompass the effect of different sensor positions with the goal to increase the algorithm's robustness.

Chapter 2

Related Work

Radar-based heart sound detection has been recently introduced by Will et al. in 2018 [Wil18]. Before this work, continous-wave (CW) radar and microwaves were already used to detect heart beats [JC92; Mat00; Iwa21]. However, these approaches tried to recover the HR by looking at the typical frequency of a beating heart lying in the 0.7...3.3 Hz frequency band [Keb20]. In contrast, the surface vibration caused by the closure of the heart valves is expected in the frequency range of 16 to 80Hz [Hol74]. In the following, different approaches that have been developed so far to retrieve heart sounds from radar measurements since [Wil18] will be presented. Further, it is looked at work beyond the proof-of-concept that tried to assess and improve the developed approaches in presence of typical challenges (optimal position of radar antenna, RLBM). Lastly, a summary of datasets, that have been published on the topic so far is given.

2.1 Heart Sound Detection Algorithms

The next two paragraphs will present two algorithms retrieving heart sounds from radar measurements which have been published so far. They differ mostly in the mathematical model used to detect the heart sounds from the processed radar data. One is based on a HSMM [Wil18] and the other one on a biLSTM [Shi19], respectively. The radar data used to train those two different models is, however, retrieved in a similar way. A Six-Port interferometer is used to retrieve relative distance changes to the reflecting obstacle from the phase shift between the transmitted and received radio waves [Koe16]. The measured distance changes are then bandpass-filtered in the 10[Shi19]/16[Wil18]...80 Hz band to exclude all movement sources apart from the heart valve closure.

Logistic Regression Hidden Semi-Markov Model The approach developed by Will et al. to now segment the heart sounds from these relative distance changes was inspired by work of Springer et al. from 2016 [Spr16]. The latter have developed a HSMM for the automatic segmentation of the FHSs in a PCG recording and Will et al. used the same approach to do so with the radar recording described above instead of the PCG. The first essential component of this probabilistic model for heart sound segmentation is an estimate of a probability density function describing the time expected to remain in each state of the HSMM, i.e. the incorporation of prior knowledge. These states are: 1) the S_1 sound, 2) the systolic period between S_1 and S_2 sound, 3) the S_2 sound, and 4) the diastolic period between S_2 and S_1. The expected HR-dependant state durations have been determined in previous work with a similar approach [Sch10]. From sequential observations of the relative distance changes, state history and prior knowledge about the expected state durations, the HSMM then infers the likelihood to be in each of the possible states.
The authors found that the F-scores for the predicted FHS from the radar are close to the gold standard predictions made from PCG data (92.22 $\pm 2.07\%$ vs. $94.15 \pm 1.61\%$). This result, paired with high a correlation of around 80% between radar and PCG data proves that radar based systems can be an alternative to PCG for measuring heart sounds. Above that, when looking at the predicted IBI, the authors showed that an IBI computed from the radar-based heart sound predictions yields a significantly lower error to a reference ECG than doing so from state of the art heartbeat detection algorithms like [Wil17]. Computed from the predicted heart sounds a root mean squared error (RMSE) for the IBI of 44.2 ms is achieved in contrast to 144.9 ms for the state of the art heartbeat variant.

Bi-Directional Long Short-Term Memory Another approach to detect heart sounds using a biLSTM was developed by Shi et al. in 2019 [Shi19]. The biggest difference, besides the used model architecture, is that the biLSTM does not require any prior information about the expected time to remain in each of the four previously mentioned states (S_1, systole, S_2, diastole). The long short-term memory (LSTM) architecture is a special variant of a recurrent neural network (RNN) and is used to model non-linear long-term dependencies in the data [Hoc97]. To perform heart sound segmentation using the biLSTM features from the same bandpass-filtered distance information as was input for the HSMM were computed. The three features they computed and used as input for their model were the homomorphic envelogram (HoEnv), the Hilbert envelope (HiEnv) and the power spectral density envelope (PSDEnv). Further, they have varied some hyperparamters of the model, especially the number of hidden units of the biLSTM.
With the best performing model, having 200 biLSTM units, they managed to achieve a F1-score on the first heart sound, S_1, of 95.8$\%$ and a F1-score of 87.7$\%$ for both heart sounds combined.

2.2 Influence of Sensor Positions and Body Movements on Heart Sound Predictions

The high performances of the two methods described above have been achieved for measurements where the person under test (PUT) was lying in bed [Shi19] or seated comfortably in an arm chair breathing at leisure [Wil18]. In a real-life scenario as it can, e.g., be expected in a infant care center, those scenarios are unrealistic. In fact, artifacts from speech and especially RLBM have to be taken into account. For this reason Herzer et al. have conducted a study in which they evaluated the performance of the HSMM algorithm for different sensor positions and in the presence of different movement protocols [Her22]. The evaluated sensor positions were lower pectoral, upper pectoral and dorsal. The performed movements were head movements, arm movements, medial-lateral (ML) torso movements and posterior-anterior (PA) torso movements. The comparison with ECG reference data showed that the dorsal sensor position yields the best results (i.e. the smallest mean absolute error (MAE) of the predicted instantaneous HR) for the baseline measurements, during head and arm movements and for moderate ML torso movements. The lower pectoral sensor position performed with the least error across sensor positions for intense ML movements and moderate as well as intense PA movements. Further, the authors did not find any correlation between the observed MAE and movement intensity.

2.3 Published Datasets

There are, so far, three published radar-recorded datasets of human vital signs. The first one curated by Shi et al. includes synchronised data which are acquired using a Six-Port-based radar system, a digital stethoscope, an ECG, and a respiration sensor [Shi20]. The data was recorded from different measurement positions (at the carotid, the back, and several frontal positions on the thorax) as well as in different scenarios such as breath-holding or post-exercise. The second dataset by Schellenberger et al. contains, besides Six-Port-based radar measurements, ECG, impedance cardiogram (ICG) and blood pressure recordings [Sch20]. The measurement protocol also included different scenarios aiming to trigger hemodynamics and the ANS of the subjects (e.g. breath-holding). Contrary to the Six-port technology, another dataset published contains frequency-modulated continuous wave (FMCW) radar-recorded vital signs for 50 children aged less than 13 years together with heart rate and respiration rate reference signals [Yoo21].

Chapter 3

Methods

In order to evaluate the performance of new technological sensing approaches for stress-related health parameters, such as respiratory rate, HR or the pre-ejection-period (PEP) against gold-standard methods from clinical practice a study was conducted at the *Machine Learning and Data Analytics Lab* from December 2022 to May 2023. The participants were asked to perform the Trier social stress test (TSST) and friendly Trier social stress test (f-TSST) in randomized order on two consecutive days. Simultaneously various sensors collected reference data as well as data that will be checked for replacement suitability of the gold-standard methods. This work was focuses on the collected radar and the ECG data. The following heart sound detection includes steps to prepare input and label data to then train a biLSTM to be able to predict heart sound occurrences from radar measurements. Finally, different hyperparameters of the model have been optimized within a grid search and evaluated towards metrics like the popular F1-score which will be presented at the end of this section.

3.1 Study Design

Since this work is part of a bigger research project focusing strongly on stress responses and their measurement, the relevant data of this work was acquired during the TSST [Kir93]. It is the gold standard method for psychological stress induction and does so through the execution of an interview followed by an arithmetic task [All17]. Previous studies have shown that the TSST reliably induces a strong increase in the stress hormone cortisol in 70-80% of study participants [Dic04]. In addition, a modified version of the f-TSST [Wie13] was carried out to obtain a control version of the experiment randomly the day before or after the TSST. Contrary to the published version of the f-TSST, in this study it included an arithmetic task as well. In both versions, the TSST as well as the f-TSST, the participant was randomly either sitting or standing while conducting the experiment.

3.1.1 Subjects

For the study, 45 participants have been recruited. In this work only the ones who have been sitting were included because the movement artifacts are expected to be lower. Later on, one of the sitting participants had to be excluded from the heart sound detection because of a missing synchronisation between radar and ECG. The demographic and anthropometric data of the participants, as utilized in the study, are listed in Table 3.1. Participants were recruited via mail distribution lists, social media, flyer notices, and in person. Eligibility for the study was assessed in advance using a screening questionnaire. In the process, those who met at least one of the following conditions were excluded: age below 18 or above 50 years, non-German native language, a BMI lower than 18 or higher than 30, physical or mental illnesses of any kind, medication intake, smoking, drug use, adiposity, and experience with a comparable stress test. As compensation, the participants could choose between 50 € or 5 Versuchspersonenstunden, if they were psychology students.

Table 3.1: Demographic and anthropometric data of the participants; Mean ± SD

	Age [years]	**Height** [cm]	**Weight** [kg]	**BMI** [kgm^{-2}]
Male	23.8 ± 3.03	180.2 ± 6.62	75.57 ± 6.68	23.31 ± 2.21
Female	23 ± 2.56	169.82 ± 4.53	60.52 ± 6.5	20.92 ± 1.39
Total	23.38 ± 2.82	174.76 ± 7.65	67.69 ± 9.99	22.06 ± 2.18

3.1.2 Experimental Protocol

Trier Social Stress Test (TSST)

The TSST follows a strict protocol with very low flexibility. The first part is a short preparation phase where the participants filled out a questionnaire and could take notes about potential key qualifications regarding their personality for the following job interview. After the preparation two tasks are following: Firstly, a simulated job interview where the participant is instructed by a jury dressed in lab coats to tell them about his or her personality. The jury always consists of one male and one female member of which is claimed that they are experts in the reading and analysis of non-verbal body gestures and facial expressions. The very little allowed interactions of the jury with the participant are almost fully scripted. To induce stress only the jury member with opposite sex (the active member) is responding to and instructing the participant. Also the participant is continuously reminded to keep eye contact with the active jury member. The second part of the test is a mental arithmetic exercise. The task is to repeatedly subtract 17 from 2043 as quickly and accurately as possible. As soon as an error is made the active jury member will interrupt and ask the participant to start again at 2043. During the experiment several short pauses are provided. The schedule of the experiment is shown in Figure 3.1.

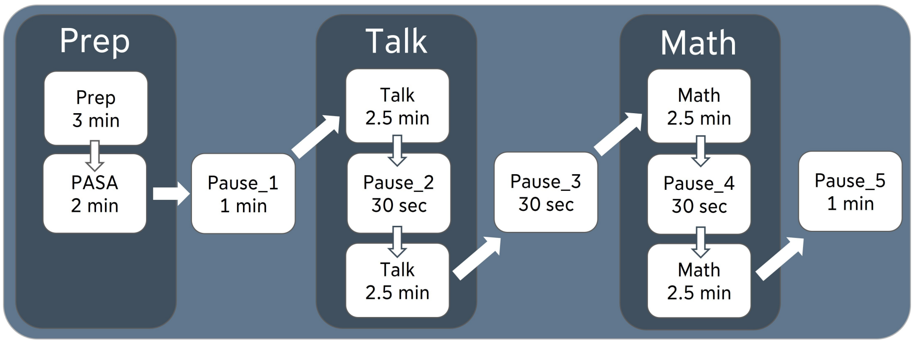


Figure 3.1: Schedule of (f-)TSST (Source: [Stü23])

Friendly Trier Social Stress Test (f-TSST)

The friendly control version of the TSST, the f-TSST, follows a very similar timeline. However, the social atmosphere was made much more comfortable during the experiment. The jury wore casual clothes and both members interacted alternately with the participant. Generally the jury showed interest and affirmation towards what the participant explained and tried to avoid any unpleasant situation during the course of the experiment. Additionally, the math task was simpler as it changed to alternately add 10 and 20 and in the event of an error the participant was allowed to continue with the last correct number.

3.2 Measurements

Several measurements were taken during the TSST and f-TSST, two of which were used for further analysis. These were a six-port interferometer for precise radar sensing and a state-of-the-art ECG. The following two sections will give an overview about these techniques.

3.2.1 Electrocardiogram

The heart is a muscular organ and its contraction and relaxation is controlled by the presence of depolarizing and re-polarizing muscle cells (called muscle fibers). The current state of polarization across the whole heart translates into an electromagnetic force, current, or vector with both magnitude and direction when measured between two electrodes surrounding the heart. The fundamental principle behind recording an ECG is that when a depolarization current travels over the heart muscle fibers towards or away from an electrode, it gets recorded either as a positive deflection or a negative deflection depending on the polarization of the measuring electrodes [Sat23]. The electrodes are placed on the surface of the skin and there are twelve configurations, the so-called leads, to place them depending on which direction you want to measure depolarization in [AL-15]. The most common configurations include limb leads, which are placed on the arms and legs to record frontal plane electrical activity. Chest leads, also known as precordial leads, are positioned on the chest to capture horizontal plane electrical signals and provide a closer look at the heart's anterior surface. Additionally, augmented limb leads (aVR, aVL, aVF) offer insights into specific cardiac areas and are derived from the limb lead electrodes [Sam15]. For this study the lead II configuration has been used. It measures the electrical activity of the heart by recording the voltage difference between the positive right arm and negative left leg electrodes. The recording vector goes from the right arm (positive) to the left leg (negative), creating a vector that points downward and to the left [Dub00]. When the heart's electrical depolarization (contraction) and repolarization (relaxation) occur in this direction, it generates characteristic waves on the ECG:

- The P wave represents atrial depolarization (atrial contraction).
- The QRS complex represents ventricular depolarization (ventricular contraction).
- The T wave represents ventricular repolarization (ventricular relaxation).

Since it is known at which points in the ECG recording the heart valves close, this measurement can be used to generate labels for heart sound detection later on.
The ECG data were acquired using the Biopac MP160 system and the AcqKnowledge software package [BIO23]. The sampling rate was set to 1000 Hz.

3.2.2 Radar

With the beginning of the first pause (see Figure 3.1) radar measurements were acquired using two or four antennae depending on whether the participant was sitting or standing. The antennae were both receiving (RX) and transmitting (TX). When sitting one antenna was focused on the lower pectoral and another one on the mid dorsal body region. In the standing scenario, there was one antenna again focusing the lower pectoral region and additionally three antennae focusing on the dorsal region at different heights (see Figure 3.2). For each antenna the radar signal has been generated using a six-port interferometer developed by Koelpin and colleagues [Koe16]. The six-port technology has been strongly promoted by Glenn F. Engen and Cletus A. Hoer who have made key publications in this field in the 1970s [Eng77]. Here, the six-port network is fed by two input signals at ports P_1 and P_2, one is the original transmitted signal and the other one the received signal which was scattered back (e.g. at the body surface) to the antenna. The two inputs are superimposed within the six-port network with static relative phase shifts of $n \cdot \pi/2$ ($n \in \{0, 1, 2, 3\}$) amongst each other. The resulting four output signals $B_{3...6}$ can be observed at the remaining four ports (see Figure 3.3). Since the relative phase shifts between the four output signals are multiples of π/2, the baseband signals can be transformed into a complex representation $\underline{\mathbf{Z}}$.

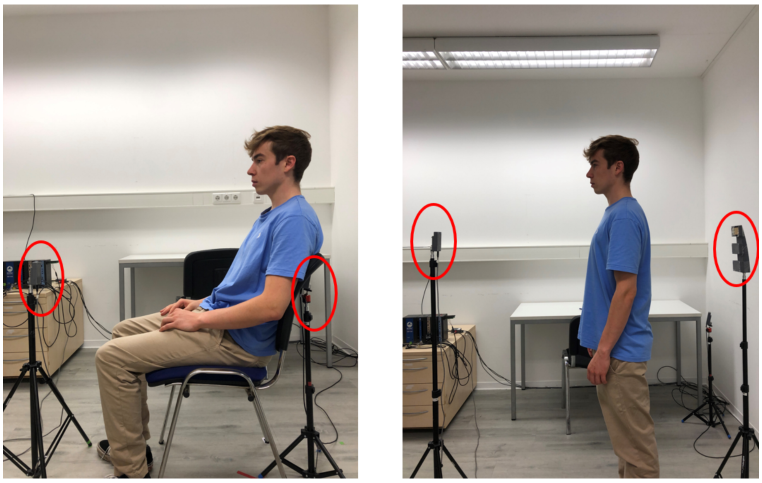


Figure 3.2: Positioning of radar antennae for sitting and standing position.

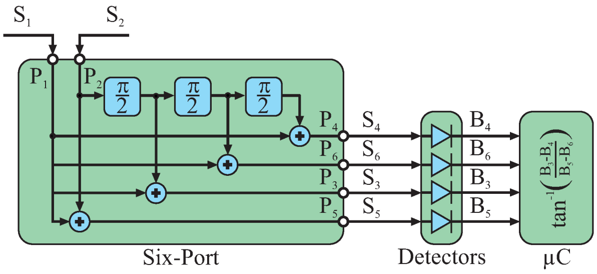


Figure 3.3: Six-Port Interferometer (Source: [Koe16])

The in-phase, as well as the quadrature component are two differential pairs, which are orthogonal to each other:

$$\underline{\mathbf{Z}} = I + jQ = (B_5 – B_6) + j(B_3 – B_4). \tag{3.1}$$

The argument of the complex expression $\underline{\mathbf{Z}}$ represents the relative phase shift between the transmitted and received electromagnetic wave:

$$\Delta\sigma = arg\{\underline{\mathbf{Z}}\} = arg\{(B_5 – B_6) + j(B_3 – B_4)\} \tag{3.2}$$

This linear phase shift can ideally be represented as movement along the unit circle, directly proportional to the target's relative distance changes. Prior work has used these distance changes as model input, but this approach will not calculate them from the raw radar (i.e., I and Q, see Equation3.1). Rather, it will use different features computed from the raw radar data, which was acquired at a sampling rate of 1952.125 Hz.

3.3 Heart Sound Detection

In this section, the complete processing pipeline is examined, including several steps common when training and evaluating machine learning models, or in this case, neural networks. Initially, the raw data underwent cleaning and preprocessing to eliminate artifacts and missing data, and to prepare it for further processing. Next, models were fed with features generated from this cleaned data, along with corresponding labels for the training phase. Following Shi and colleagues' research [Shi19], a biLSTM neural network model was implemented to segment heart sounds from radar features. The pipeline underwent optimization to enhance the model's performance. A grid search tested various combinations of parameters and hyperparameters. The evaluation of each trained model was conducted by calculating several performance metrics, which are presented at the end.

The complete pipeline, from data preparation to evaluation, has been realized using classes from the open source library *Tiny Pipelines for Complex Problems* (TPCP) [Küd23]. The library provides a higher level infrastructure for algorithm development when solutions should stay maintainable and flexible although relying on many different software packages. Further, the code realizing the following pipeline has been contributed to the repositories *empkins-io*[Erl23a] and *empkins-micro*[Erl23b]. Both libraries are developed for research purposes at the *EmpkinS collaborative research center* at the Friedrich-Alexander Universität Erlangen-Nürnberg (FAU). *Empkins-Io* is the input/output library giving access to all kinds of data relevant in studies affiliated with the research center. *Empkins-Micro* is offering functionality to analyze data that has been generated in the specific study described above (see section 3.1) that is also of concern in this work.

3.3.1 Data Preparation

The data preparation includes the synchronization of relevant raw data (i.e. ECG and radar), the aligning of sampling rates, filtering in the frequency domain and finally generating features and labels.

Preprocessing

To access the ECG and radar data, the first step was to contribute the necessary functionality to the *empkins-io* library. The loading of the ECG data as part of the Biopac recordings was realized using the library *BioPsyKit* [Ric21]. Loading radar data has already been implemented within the *empkins-io* library before. The next step was to synchronize the raw radar and ECG data. To enable this, the Biopac as well as the radar framework were connected to a *Synchronization Board (SyncBoard)*. After the recording was started by the jury, they manually added a peak to an extra synchronization channel present in both radar and biopac recordings via a graphical user interface (GUI). The SyncBoard and its accompanying GUI have been developed at the *Machine Learning and Data Analytics Lab*. When cleaning the data, two participants had to be excluded from the analysis due to missing synchronization peaks in the synchronization channel of the radar. After synchronizing, the raw radar and ECG were cut to start at the beginning of the first pause of the experiment and end together with the last pause (see Figure 3.1). Only in this window meaningful radar and ECG data were recorded. Finally, the raw radar was resampled to a sampling rate of 1000 Hz, resulting in synchronized sample pairs of ECG and raw radar. The resampling has been done using the *resampy* library which implements the band-limited sinc interpolation method for sampling rate conversion as described in [Smi02]. In order not to repeat the process of synchronizing and sample rate conversion whenever raw data is accessed again later, a notebook (for use on personal computer) and script (for use on high performance cluster (HPC)) have been created that save the resulting data as pickle files that, if available, are loaded by the synchronizing and sample rate aligning property. To prepare the raw radar for feature generation, it was highpass-filtered using a fifth order Butterworth-filter with a cutoff frequency of 0.4 Hz. Thus, the DC offset was eliminated the measurements as it is not relevant for detecting relative changes in radar signals. The filter implementation uses the popular *scipy* library [Vir20].

Feature Generation

To train and make predictions using a biLSTM, sequential data is necessary. Generally speaking, as input several sequential features will be generated for a fixed time window. The pipeline has been designed to easily allow the use of modified features or add new ones in the future. In this work four features have been used. The first two are the highpass-filtered radar in-phase (Q) and quadrature (I) components (see Equation 3.1) that were output of the preprocessing.

Additionally, the phase shift angle between the transmitted and received radio waves was calculated from I and Q as shown in Equation 3.2. Lastly the magnitude signal was computed from the highpass-filtered I and Q components and bandpass-filtered it with a lowpass cutoff frequency of 15 Hz and a highpass cutoff frequency of 80 Hz, similar to previous work [Wil18; Shi19], as the heart sounds are expected in this frequency band. Similar to Shi and colleagues, an envelope signal of the Hilbert transform was computed from this bandpass-filtered power signal. In this analysis the envelope was calculated by applying a moving average with window length 100 (i.e. 100 ms).

To recreate the proof-of-concept for this approach, the described envelope feature is shown together with the synchronized ECG measurement in Figure 3.4 for different intervals during the experiment. From the figure one can see that during the pauses of the experiment the amplitude of the Hilbert envelope is about two orders of magnitude smaller than during the phases where the participant is actively involved. When looking at the enlarged interval during Pause 1, clear matches between the R-peaks as well as T-wave ends in the ECG and peaks in the Hilbert envelopes are visible. When in the active stages of the experiment, those matches are not clear anymore and peaks in the Hilbert envelope are likely due to motion artifacts. All four features have been downsampled by a factor 10 such that they are represented with a sampling rate of 100 Hz. Doing so speeds up processing by a factor 10 while not loosing frequency information due to the computation of the envelope. Before fed into the network the features were cut into windows of 400 samples (i.e. 4 seconds) overlapping each other by 95%. Both the overlap and the window length can be changed easily for the pipeline. They have been chosen like this to, firstly, ensure that a few heart sounds are sure to occur in each input sample and, secondly, to generate a sufficient amount of training data. Every training sample was additionally normalized using the min-max normalization to fit the interval [0,1]. It was found that when not doing so and normalizing the whole input dataset together the model will not even slightly converge during training. The reason for that is the much higher amplitude of motion artifacts in the radar compared to the heart sounds (see Figure 3.4). Without adapted normalization, it would be very difficult for the network to learn from low-noise data samples.

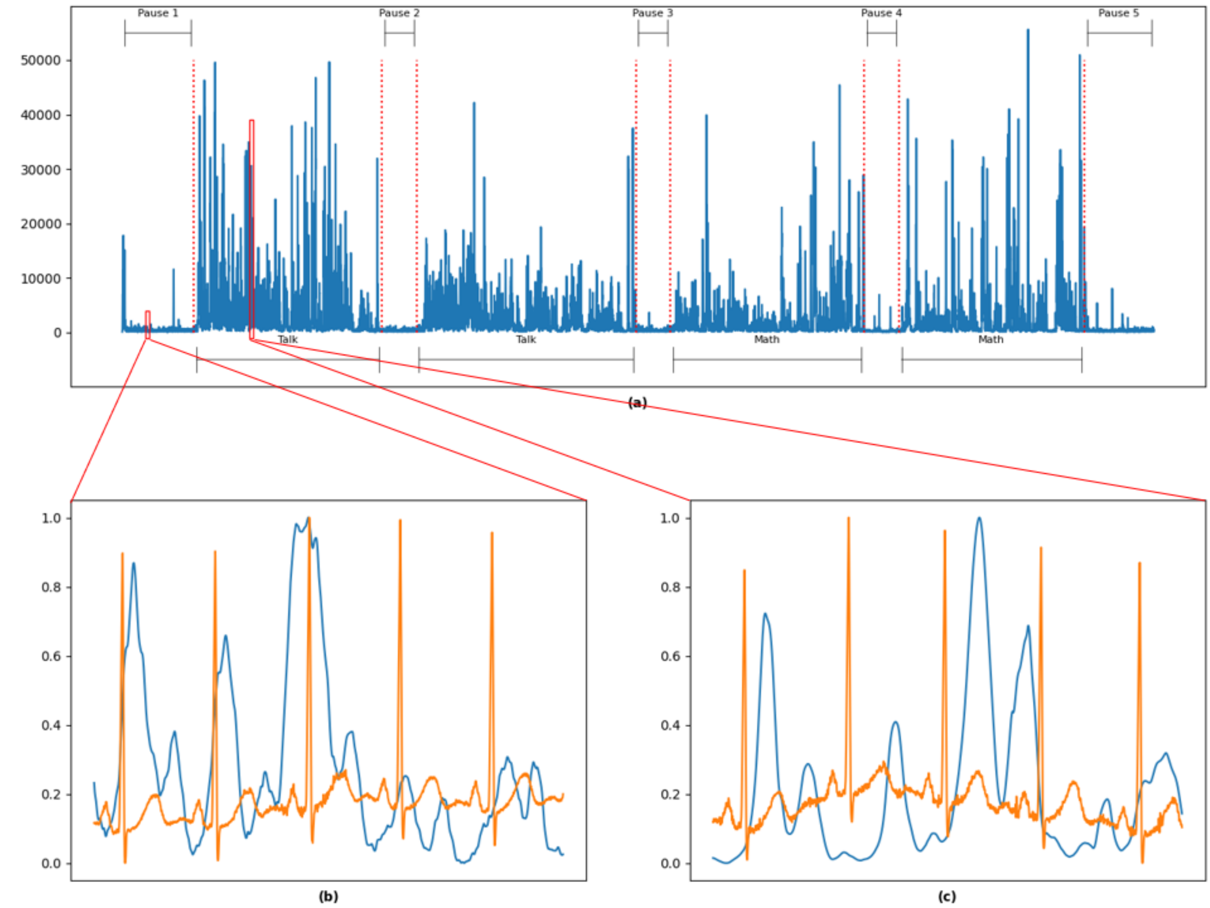


Figure 3.4: (a): Radar envelope for one participant and a complete experiment; (b): Cut out window of radar envelope normalized within window (blue) and synchronized ECG during Pause 1; Cut out window of radar envelope normalized within window (blue) and synchronized ECG during Talk.

Label Generation

The used labels are solely based on the R-peaks in the ECG signal. As described in chapter 1, the R-peaks are occurring approximately with the closing of the atrioventricular valves (i.e. the first heart sound). Besides, the R-peaks are the most prominent feature in the ECG signal and are relatively easy to detect and therefore yield the most reliable labels. As a first step the ECG was downsampled by a factor 10 as well and hence represented with a sampling rate of 100 Hz. Afterwards the R-peaks were detected in the ECG signal using the automated pipeline for preprocessing an ECG signal, called *ecg_process*, developed as part of the *neurokit2* library [Mak21], a python toolbox for neurophysiological signal processing. The *ecg_process* function produces a list of zeros and ones corresponding to detected R-peaks. To enhance label smoothness for better model learning, a Gaussian kernel of default length 400 with zero mean and standard deviation of 6 is applied to the list of 1's and 0's, resulting in a R-peak probability distribution. The resulting signal of Gaussians centered at detected R-peaks is used to generate the labels by cutting it into windows that have the same length of 400 samples and an overlap of 95% as the input features.

3.3.2 Detection Model

After investigating input and label generation, the subsequent section will present an overview of the neural network utilized in the pipeline. Building on the work of Shi and colleagues summarized in chapter 2.1, the developed neural network also incorporates a layer consisting of biLSTM units [Shi19]. Being bidirectional, this layer has the advantage to incorporate input in forward and backward direction. Effectively this takes into account that sequential data carries information in forward as well as backward direction. Besides, there is one additional layer consisting of mono-directional LSTM units to increase complexity of the model and finally a dense, fully connected layer. The dense layer weighs the output of the LSTM layer to arrive at a final embedding matching the dimensions of the labels. In between the biLSTM, LSTM and dense layer there are dropout layers to prevent the model from overfitting [Sri14]. A dropout layer is assigned a certain dropout probability p. During training only the output of $p \cdot num\_units$ that went into the dropout will be passed to the next layer for a single batch. Consequently, also only the weights of the units not dropped out will be updated after each batch. During inference after training, the dropout becomes 0 and hence, all units are used. A dropout has been used due to the strong overlap, and therefore repetitiveness, of training data described in section 3.3.1. From the provided shape of the input data (default is (400, 4), see section 3.3.1), the input shape of the network as well as the number of units in the final dense layer are automatically determined. Many other hyperparameters, however, can be further optimized. Those include, the number of units in the biLSTM and LSTM layer, the dropout rate of the two dropout layers, the learning rate, the batch size and the number of epochs spent training the model.

Long Short-Term Memory

Since the LSTM architecture plays a crucial rule for this development pipeline, its functionality will now be summarized. LSTMs were introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997 [Hoc97]. They are a special kind of a RNN that does not only have a short term memory but also a long term memory which tackles the problem of vanishing/exploding gradients of vanilla RNNs [Ben94]. This problem makes it impossible to learn either long term or short term dependencies from the data. Each LSTM unit consists of a short term memory, a long term memory and three gates:

1. The *Forget Gate* which controls how much % of the current long term memory is remembered and used from this unit onward.
2. The *Input Gate* which controls how the long term memory is updated in this unit.
3. The *Output Gate* yielding the output of the unit as a combination of short and long term memory. The Output is also used as the short term memory for the following LSTM unit.

Each LSTM unit has 8 weights and 4 biases to modify during training. In the case of the biLSTM the weights and biases double due to the number of units doubling. When using more than one feature, this also increases the number of weights and one arrives at the following formula for computing the number of weights and biases in a LSTM:

$$num\_parameters = 4 \cdot [h \cdot (h + e) + h] \tag{3.3}$$

With h being the number of LSTM units and e the number of features. From that, it is easy to see that LSTM architectures are very computationally expensive to train.

Additional to the layer of LSTM units processing sequential data in forward direction, a biLSTM consists of a second LSTM layer processing the same sequential data in backward direction. This respects the fact that information about the following member in a sequential chain can flow both forward and backward.

3.3.3 Optimization

For the optimization and testing of the developed pipeline the data of all sitting participants was included, if complete. The seated position is expected to induce weaker motion artifacts and thus a better signal-to-noise ratio to start with. In total, following the described default data preparation (see section 3.3.1) and a batch size of 128, 1100 batches are available for training and validation of the model. The number of training epochs was set to 25 and the learning rate to 0.001 as this ensured the model loss to converge sufficiently during training. In order to optimize the model structure, we conducted a grid search utilizing k-fold cross-validation with non-overlapping participant groups. The parameter grid used is as follows:

- Number of biLSTM Units: $\{64, 128\}$

- Number of LSTM Units: $\{128, 256\}$
- First Dropout Rate: $\{0.3, 0.6\}$
- Second Dropout Rate: $\{0.3, 0.6\}$

Thus, the study evaluated 16 distinct configurations in relation to three defined metrics, as outlined in the subsequent section 3.3.4. The performed grid-search used the *GroupKFold* method from the scikit-learn library for applying a 5-fold during optimization. The groups were chosen to be on a per-participant level. In this way, the model could not learn to make predictions for a participant in one of the two study conditions and be tested in the other.
Since the grid search required five trainings for a single parameter combination, it necessitated the use of powerful computing resources. For this purpose, scripts were developed that allow for access to the *TinyGPU* HPC cluster, maintained and provided by the *Erlangen National High Performance Computing Center*. All optimization has been performed on a single *NVIDIA GeForce RTX3080* graphics processing unit (GPU).
The dorsal antenna was shown to be one of the most robust against motion in the study of Herzer and colleagues [Her22]. Because of that, it was used during the grid search. To evaluate the optimized pipeline a second antenna focusing the front of the subject will be compared to the dorsal antenna in the results section.

3.3.4 Validation

To evaluate the effectiveness of the trained models, various metrics were calculated. Two of the three computed metrics are based on the IBI of the predicted R-peaks. The predicted R-peaks were detected using the*find_peaks* function from the *scipy* library. The peak detection process has been constrained to identify peaks solely when they were separated by a sampling interval equal to or greater than a maximum heart rate of 180 beats per minute (bpm). Furthermore, based on empirical testing of different values a minimum prominence of 0.15 was determined to be a suitable balance between detecting enough peaks while only counting the relevant ones. Given the location of identified R-peaks, for every two consecutive pair of R-peaks, the instantaneous heart rate can be calculated from the sample difference separating the two:

$$pairwise\_heart\_rate = 60 * \frac{sampling\_rate}{peak\_sample\_distance}. \tag{3.4}$$

The next step was to interpolate over the computed sequence of pairwise heart rates and sample the resulting interpolation with a sampling frequency of 1 Hz. If done for the labels as well as the resulting prediction, the MAE of the predicted instantaneous heart rate can be determined. Additionally, a moving average over the heart rate prediction and ground truth with different window sizes has been calculated. Between the two, the Pearson correlation coefficient [Fre07] can be computed as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{3.5}$$

Finally, the performance of the trained model can be evaluated by calculating the F1-score with respect to the detected R-peak locations. True positives (correctly detected R-peaks) can be counted for different tolerance windows, which are centered around the ground truth R-peak locations and have a range of $\pm 50ms$, $\pm 100ms$, and $\pm 200ms$. The F1-score is then computed using the following formula:

$$precision = \frac{\#true\_positives}{\#rpeaks\_predicted}, \tag{3.6}$$

$$recall = \frac{\#true\_positives}{\#rpeaks\_groundtruth}, \tag{3.7}$$

$$F1 - Score = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{3.8}$$

To detect the grid search optimization, the different pipeline configurations have been ranked according to the achieved $F1 - Score$ on the test split.

Chapter 4

Results

In the following, we will have a look at the results of the developed approach. First the outcome of the described grid search will be analysed. Afterwards, the best performing model configuration will be evaluated in more depth.

4.1 Grid Search

Table 4.1 shows the results of the performed grid search over the parameter grid of the 16 different model configurations using the antenna focusing the back of the participant. The MAE of the predicted instantaneous HR over all experimental phases is in the range of 21.57 to 23.46 bpm. The mean F1-score for a tolerance window of ± 100 ms around the ground truth R-peak is in the range of 0.435 to 0.481. All obtained results are very close to each other and there is no trend in favor of any particular combination or subset of tested hyperparameters. However, the configuration that performed best in cross-validation with respect to the mean F1 score across all folds was selected for further evaluation. This is 64 biLSTM units, 128 LSTM units and a dropout of 0.3 for both dropout layers. The selected pipeline was rebuilt with the chosen configuration and retrained for 50 epochs on all available data except for a single participant. This enabled the use of the largest possible quantity of training samples while avoiding evaluation of the trained model on data from a participant it has previously encountered. The data from the ”left out” participant was then used to evaluate the resulting model. This leave-one-out cross-validation procedure was repeated for all participants, and finally the evaluation results were combined.

Table 4.1: Results of Grid Search; Mean ± SD of each score for 5-fold cross-validation. Trained pipelines have been ranked according to achieved F1-score.

First Dropout	Second Dropout	Mono-LSTM Units	biLSTM Units	MAE Instantaneous HR	F1-Score (100 ms)
0.3	0.3	64	128	21.82±3.66	0.481±0.046
0.3	0.3	64	256	22.4±2.96	0.472±0.045
0.3	0.3	128	128	23.13±2.53	0.469±0.042
0.3	0.3	128	256	21.79±3.14	0.474±0.044
0.3	0.6	64	128	23.46±4.75	0.443±0.08
0.3	0.6	64	256	22.13±2.77	0.458±0.047
0.3	0.6	128	128	22.55±3.07	0.435±0.072
0.3	0.6	128	256	22.28±3.26	0.476±0.049
0.6	0.3	64	128	22.78±3.08	0.479±0.044
0.6	0.3	64	256	23.24±3.02	0.47±0.046
0.6	0.3	128	128	22.06±3.59	0.48±0.048
0.6	0.3	128	256	21.57±3.49	0.473±0.048
0.6	0.6	64	128	23.12±2.72	0.476±0.045
0.6	0.6	64	256	22.97±3.17	0.475±0.038
0.6	0.6	128	128	23.39±2.78	0.45±0.071
0.6	0.6	128	256	22.52±2.86	0.476±0.048

4.2 Evaluation of Best Performing Model Configuration

Only the dorsal antenna has been used to run the grid search and find the most promising parameter configuration. However, To gain insight into the influence of sensor positions and the robustness of the optimized pipeline configuration, we compared the two different sensor positions (dorsal and frontal) for seated participants regarding the selected pipeline configuration.

Predicted Heart Rate The trained model of one exemplary pipeline from the Leave-One-Out cross-validation using the frontal sensor position was run on the data of the corresponding subject left out during training. From the resulting R-peak prediction, the instantaneous HR was calculated as described above, and the same was done for the labels. Figure 4.1 shows a plot of the predicted together with the ground truth HR after calculating a moving average with a window size of 30 seconds. From the Figure it can be inferred that the trained model is able to detect trends regarding the development of the HR over the course of an experiment. This result is mathematically underpinned by a computed Pearson correlation of 0.782 between the prediction and ground truth.

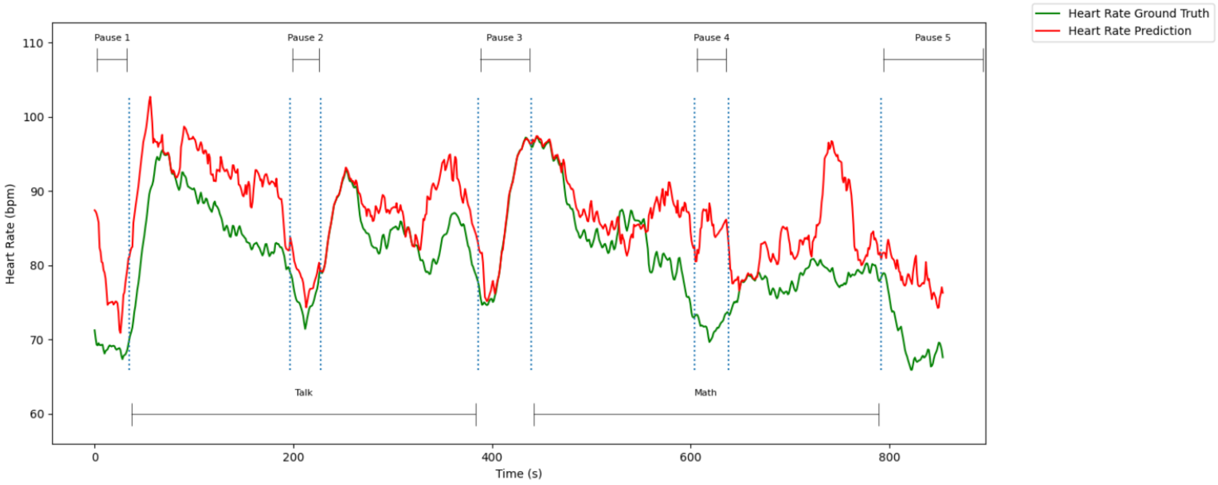


Figure 4.1: Moving average with a window size of 30 seconds computed for the beat-to-beat HR prediction and ground truth of one omitted participant during the TSST.

Absolute Error of Instantaneous HRs per Experimental Phase This evaluation took into account all resulting pipelines from the leave-one-out cross-validation. It compared the absolute error (AE) of the predicted HRs among the various experimental phases (see Figure 4.2). From the plot it can be inferred that the error of the predicted HR is the highest for the *Talk* followed by the *Math* part. These phases are the active parts of the experiment where the participant can be expected to have moved much more than during the pauses. However, there is often still a significant absolute error of around 10 bpm even when considering evaluation during pauses. Furthermore, regarding the different sensor positions used, the antenna focusing the front of the participant clearly outperforms the antenna focusing the back (see Figure 3.2 for position of antennae). This finding is consistent for all experimental phases.

Error of Predicted Instantaneous HRs With Regard to Ground Truth HR In the next evaluation it was investigated using Bland-Altman plots whether the error of the predicted HR was in any way affected by the ground truth HR for both sensor positions (see Figures 4.3 and 4.4). It was found that for low HRs the prediction generally is too high while for high HRs the prediction is too low. The model tends to predict R-peaks corresponding to a HR being at the midpoint between the two extremes of around 40 and 160 bpm. A possible explanation of this tendency will be given in the Discussion section 5. Further, although this tendency is present for both antennae positions, it is weaker for the frontal antenna. This can be inferred from the smaller standard deviation from the mean error being 56,8 for the dorsal antenna while being 48.8 for the frontal antenna.

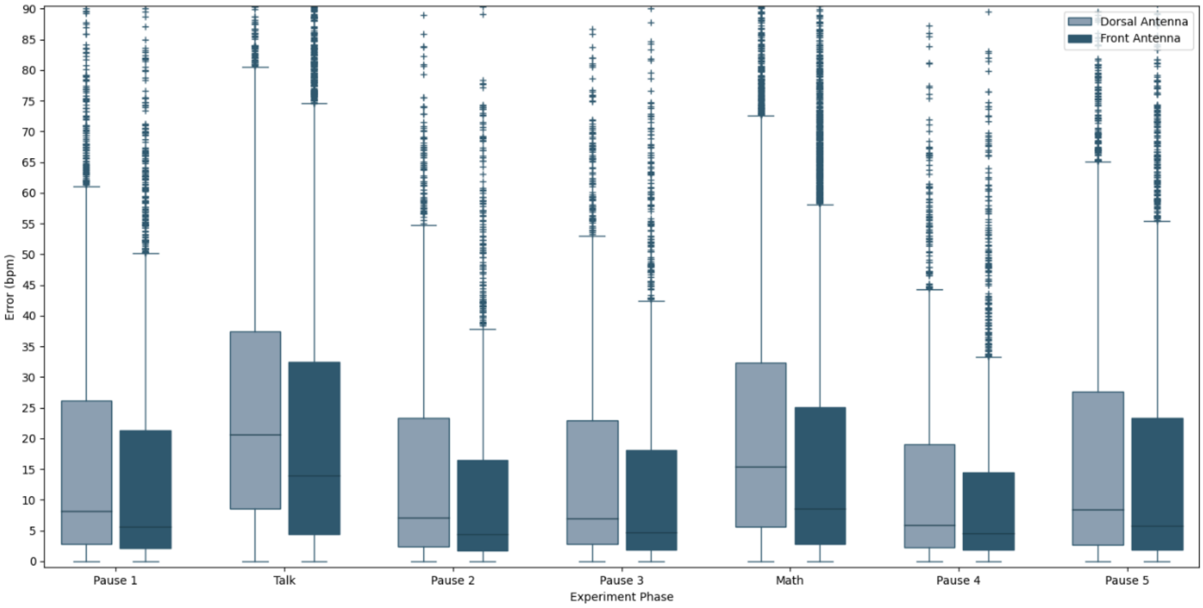


Figure 4.2: AE of the predicted instantaneous HRs per experimental phase for leave-one-out cross-validation across the entire dataset for two different sensor positions.

F1-Scores Per Experimental Phase for Different R-Peak Tolerances The final assessment analyzed the F1 scores for the prediction of R-peaks (Table 4.2) throughout the entire cross-validation and considering the used sensor position. The F1-scores were computed for each experimental phase, incorporating various tolerance windows encompassing the actual R-peaks. Similar to the results obtained from the grid search, the *Talk* and *Math* phases produced the lowest scores for both sensor positions, while better predictions were made for the data collected during pauses. Furthermore, throughout all experimental phases and tolerance windows, the F1-score was consistently higher for the frontal sensor. Especially notable is the change of the F1-score on all data with a tolerance window of 100 ms which increased from 0.473 when using the dorsal sensor to 0.6 when using the frontal sensor.

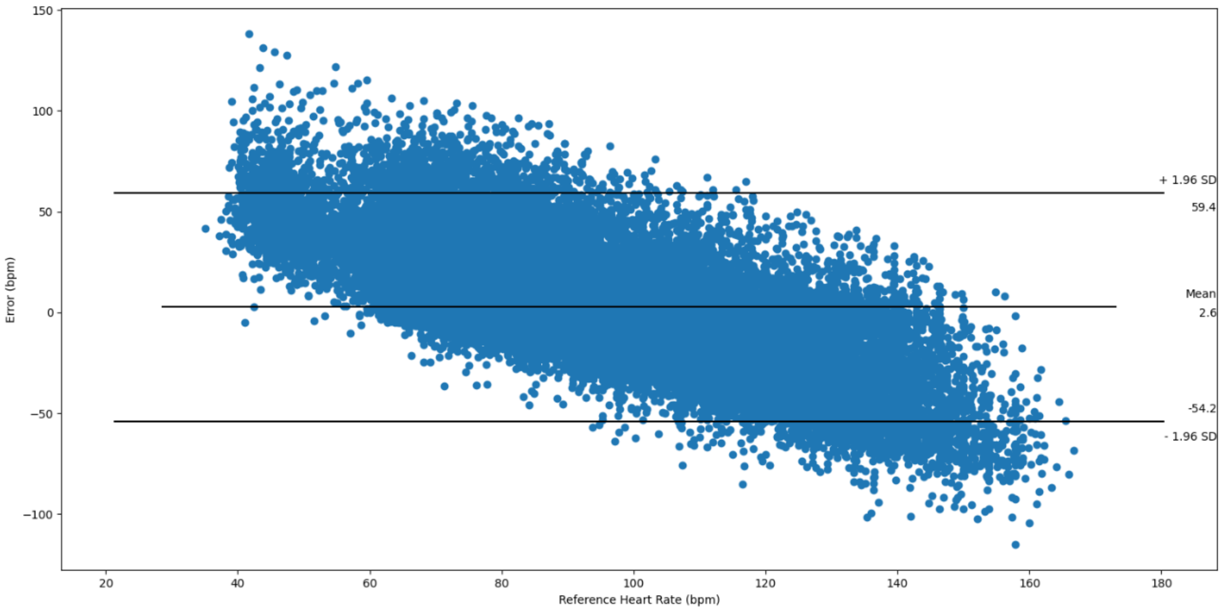


Figure 4.3: Error of the predicted instantaneous HRs with regard to the ground truth heart rate for the dorsal sensor position

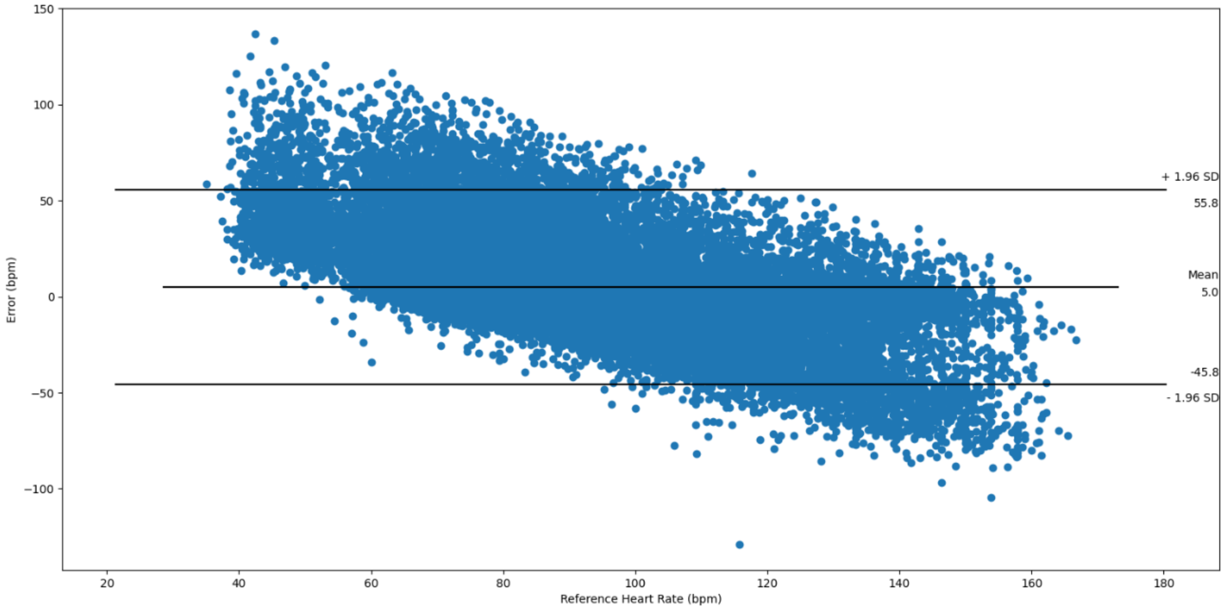


Figure 4.4: Error of the predicted instantaneous HRs with regard to the ground truth heart rate for the frontal sensor position

Table 4.2: F1 Score for R-peak prediction with different R-peak tolerances comparing two sensor positions; Mean ± SD

Experimental Phase	Dorsal Antenna			Pectoral Antenna		
	F1-Score (50 ms)	F1-Score (100 ms)	F1-Score (200 ms)	F1-Score (50 ms)	F1-Score (100 ms)	F1-Score (200 ms)
All Data	0.26 ± 0.07	0.47 ± 0.10	0.73 ± 0.10	0.41 ± 0.18	0.60 ± 0.19	0.78 ± 0.13
Pause 1	0.33 ± 0.13	0.59 ± 0.18	0.82 ± 0.13	0.406 ± 0.23	0.65 ± 0.24	0.83 ± 0.16
Talk	0.21 ± 0.06	0.40 ± 0.08	0.69 ± 0.09	0.34 ± 0.17	0.52 ± 0.18	0.74 ± 0.13
Pause 2	0.44 ± 0.21	0.67 ± 0.22	0.84 ± 0.15	0.51 ± 0.26	0.72 ± 0.25	0.85 ± 0.17
Pause 3	0.43 ± 0.18	0.67 ± 0.20	0.85 ± 0.14	0.58 ± 0.26	0.74 ± 0.23	0.85 ± 0.16
Math	0.24 ± 0.07	0.45 ± 0.12	0.72 ± 0.11	0.43 ± 0.21	0.62 ± 0.21	0.79 ± 0.14
Pause 4	0.35 ± 0.17	0.65 ± 0.22	0.84 ± 0.16	0.57 ± 0.24	0.74 ± 0.22	0.85 ± 0.15
Pause 5	0.28 ± 0.14	0.55 ± 0.18	0.78 ± 0.14	0.49 ± 0.21	0.69 ± 0.21	0.81 ± 0.16

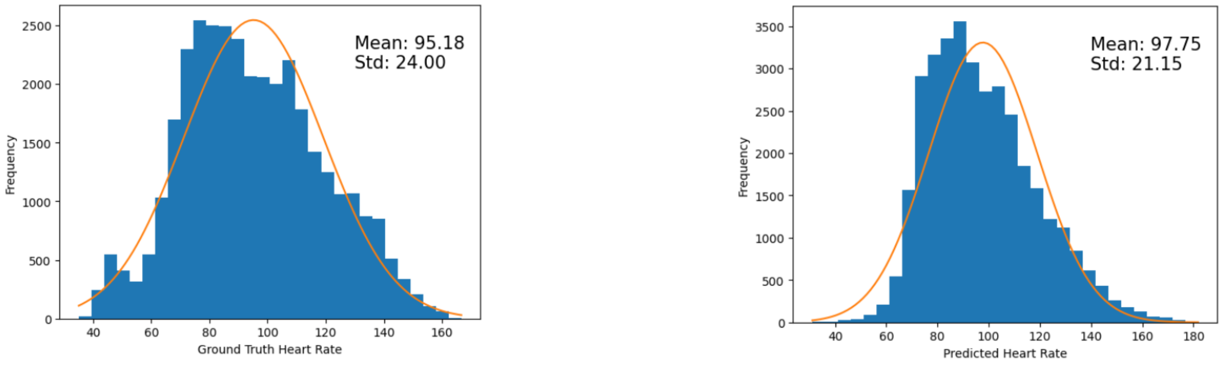
Chapter 5

Discussion

This thesis aimed to enhance the stability of radar-based detection of heart sounds against the influence of RLBM that arise exemplary during the conduction of the TSST and f-TSST. Unlike earlier works, the heart sound detection model employed in this study was trained on data that already includes motion artifacts. Therefore, the state-of-the-art model was assessed extensively for the first time with data containing motion artifacts after being trained on similar data.

Different configurations of a state-of-the-art heart sound detection model were compared to determine their accuracy in detecting the first FHS corresponding to the R-peak of the ECG signal. The study showed that tuning the dropout rates and the number of units in the LSTM and biLSTM layers did not considerably affect the accuracy of the model. Thus, the model's performance appears to be almost independent from the tested hyperparameters. However, this does not necessarily imply that it will hold for future versions of the model with differing architectures, input features or labels. It may be worthwhile in future research to evaluate the influence of these hyperparameters again and also examine alternative search space options.

To further analyze the performance of the model, the predicted heart rate has been computed from the predicted R-peak locations. Analyzing the result showed that the model favors to predict R-peaks in a frequency that corresponds to a HR of around 100 Hz (see Figures 4.3 and 4.4). For heart rates varying from this frequency, the error is quickly much higher than could be permissible for an adaption of this approach in, e.g., clinical practice. Future research in this domain will need to address this issue. A possible explanation for the strong favor of a certain HR is that the ground truth HR is focused around a similar value. With labels not evenly distributed it could be hard for the model to learn detecting heart sounds reliably in cases where the heart rate is deviating from the most common frequency. To illustrate this problem Figure 5.1a and 5.1b show how the predicted and ground truth heart rate is distributed over all data. The true HR data appears to conform to a normal distribution and the prediction closely approximates it. For most of the recorded data, motion noise interferes with the features. As a result, the model may have learned to predict R-peaks to approximate the average HR and reduce errors during training. To tackle this issue, upcoming studies may consider creating a more well-balanced dataset for training. To accomplish this aim, it is important to evenly distribute the ground truth HR data among all potential HRs. This entails the model's ability to aptly identify lower and higher HR values to reduce error and prevent it from being overly rewarded for predicting the average of the ground truth data. Further, another idea could be to use Gaussians with smaller standard deviation during label generation (see section 3.3.1). That way, in order for the loss function to converge, the model would need to predict R-peak locations more accurately. It is possible that the current default label generation allows the model excessive flexibility in forecasting R-peak locations, while insufficiently encouraging identification of the actual peaks.



(a) Label Heart Rate Distribution

(b) Predicted Heart Rate Distribution

Figure 5.1: Distribution of ground truth and predicted HRs calculated from the R-peak locations.

Further regarding the data, this study discovered that normalizing the input data all at once hinders the model architecture's convergence. Features contaminated by motion artifacts from RLBM have an amplitude two orders of magnitude higher compared to the data affected by heart sound-induced motion alone. The normalization of data with artifacts results in a significant reduction in the influence of small, artifact-free data sequences, thereby rendering them negligible during the training process. Normalizing individual input samples enabled the model to achieve convergence. Nonetheless, it failed to address the challenge of extracting heart sound information from input data with motion artifacts. In future research, the primary challenge will be to develop methods for retrieving heart sound data from radar measurements during motion-induced changes in relevant features. This study utilized windowed normalization to prevent the model from disregarding heart sound features that were less affected by motion during training. However, to train a machine or deep learning model with data containing motion, it is highly probable that data quality will need to be improved first. Thus, a combination of state-of-the-art hardware with solutions from the fields of signal processing and information theory will be necessary to differentiate signals originating from FHSs from those originating from RLBM. For instance, examining how overall performance changes when utilizing signals from multiple radar antennae could be a compelling study. That way, it may be possible to use the combined information on the direction of motion to differentiate motion related to heart sound from RLBM. In addition, together with using more uniformly distributed labels, future studies may want to include ground truth motion data. The inability to quantify motion is a significant limitation of this work. Thereby it was only possible to relate the errors across different experimental phases qualitatively to the expected movement frequency and intensity.

Finally, as a first step towards better data, the study indicates that radar measurements are more precise when taken from seated individuals with a front-focused antenna rather than a back-focused antenna. This suggests that in the presence of RLBM for seated subjects, frontal radar measurements are the most reliable. In contrast to earlier research conducted by Herzer and colleagues [Her22] which analyzed the influence of sensor positions (as described in section 2.2), the present study includes distinct variations. Primarily, the pipeline methodology differs from that of Herzer's study in that this analysis utilizes a biLSTM neural network, instead of an HSMM. Additionally, the model is not trained on stationary data, but rather on the same type of data that will be evaluated, specifically, radar data with artifacts from RLBM. Finally, the current motion is of a different nature, as it includes random and mixed movements throughout the data acquisition, rather than semi-standardized motion. As a result, the findings make a stronger statement towards realistic scenarios.

Chapter 6

Conclusion and Outlook

In the context of this bachelor thesis, a dataset for the development of radar-based heart sound detection pipelines could be generated. It consists of synchronized radar and ECG measurements with recordings of subjects in seated or standing positions from multiple sensor locations. The dataset comprises RLBM, making it suitable for testing novel approaches aimed at improving prediction performance when such motion is present. For the analysis performed in this thesis all the developed software has been contributed to the open source software packages empkins-io [Erl23a] and empkins-micro [Erl23b] and can be reused.

The robustness of state-of-the-art radar-based heart sound detection models against RLBM was assessed when trained on data that already includes such motion. This thesis demonstrates that relying solely on similar data for training does not ensure the model's robustness against RLBM. Furthermore, changing the hyperparamter configurations of the model used for prediction did not significantly enhance its performance. It is likely that the challenge of motion-robust heart sound detection will need to be solved from a data-centric rather than a model-centric point of view. Although adjusting the model's architecture or optimizing its hyperparameters are crucial steps for enhancing its final performance, the most significant opportunity for improvement may lie in obtaining and processing superior data. Analysis of the features clearly showed that FHS are lost or at least well hidden in the data under the presence of motion artifacts. Besides the features, this may also include the distribution of incorporated ground truth data. It was shown that the model learned to predict R-peak locations in a frequency corresponding approximately to the mean HR of the labels. This may minimize the overall error during training but more importantly reveals the inability of the model to detect the actual heart sounds. If future work uses uniformly distributed labels the model may be forced to generalize better to heart sounds occurring in all frequencies. Additionally, future work could combine the data gathered in this study with motion-free data from different studies. Doing so opens up possibilities for training and subsequent refinement of developed models using data with different amounts of artifacts.

A first step towards improving the quality of the data has been taken in this thesis by finding radar measurements taken from the front result in significantly lower errors than if taken from the back of a subject. In future work, it could be interesting to also combine the acquired data from multiple sensor positions or further optimize the focus spot of the antennae for optimal performance.

List of Figures

List of Tables

Bibliography

[AL-15] Hussain AL-Ziarjawey. “Heart Rate Monitoring and PQRST Detection Based on Graphical User Interface with Matlab”. en. In: *International Journal of Information and Electronics Engineering* (2015). ISSN: 20103719. DOI: 10.7763/IJIEE.2015.V5.550. URL: http://www.ijiee.org/index.php?m=content&c=index&a=show&catid=52&id=602 (visited on 09/18/2023).

[All17] Andrew P. Allen, Paul J. Kennedy, Samantha Dockray, John F. Cryan, Timothy G. Dinan, and Gerard Clarke. “The Trier Social Stress Test: Principles and practice”. en. In: *Neurobiology of Stress* 6 (Feb. 2017), pp. 113–126. ISSN: 23522895. DOI: 10.1016/j.ynstr.2016.11.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352289516300224 (visited on 01/16/2023).

[Ben94] Y. Bengio, P. Simard, and P. Frasconi. “Learning long-term dependencies with gradient descent is difficult”. In: *IEEE Transactions on Neural Networks* 5.2 (Mar. 1994). Conference Name: IEEE Transactions on Neural Networks, pp. 157–166. ISSN: 1941-0093. DOI: 10.1109/72.279181.

[BIO23] BIOPAC. *MP160 Starter Systems | BIOPAC*. en-US. 2023. URL: https://www.biopac.com/product-category/research/systems/mp150-starter-systems/ (visited on 09/18/2023).

[Bre19] Idar Johan Brekke, Lars Håland Puntervoll, Peter Bank Pedersen, John Kellett, and Mikkel Brabrand. “The value of vital sign trends in predicting and monitoring clinical deterioration: A systematic review”. In: *PLoS ONE* 14.1 (Jan. 2019), e0210875. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0210875. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6333367/ (visited on 09/26/2023).

[Dic04] Sally S. Dickerson and Margaret E. Kemeny. “Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research”. In: *Psychological Bulletin*

130.3 (2004). ISBN: 0033-2909 (Print)\n0033-2909 (Linking), pp. 355–391. ISSN: 00332909. DOI: 10.1037/0033-2909.130.3.355. URL: 10.1037/0033-2909.130.3.355.

[Dub00] Dale Dubin. *Rapid Interpretation of EKG's*. 6th. Cover Publishing Company, 2000. ISBN: 978-0912912066.

[Eng77] G.F. Engen. "The Six-Port Reflectometer: An Alternative Network Analyzer". In: *IEEE Transactions on Microwave Theory and Techniques* 25.12 (1977), pp. 1075–1080. DOI: 10.1109/TMTT.1977.1129277.

[Erl23a] EmpkinS FAU Erlangen-Nürnberg. *empkins-io*. https://github.com/empkins/empkins-io. 2023.

[Erl23b] EmpkinS FAU Erlangen-Nürnberg. *empkins-micro*. https://github.com/empkins/empkins-micro. 2023.

[Fre07] David Freedman, Robert Pisani, and Roger Purves. "Statistics (international student edition)". In: *Pisani, R. Purves, 4th edn. WW Norton & Company, New York* (2007).

[Her22] Liv Herzer, Annika Muecke, Robert Richer, Nils C. Albrecht, Markus Heyder, Katharina M. Jaeger, Veronika Koenig, Alexander Koelpin, Nicolas Rohleder, and Bjoern M. Eskofier. "Influence of Sensor Position and Body Movements on Radar-Based Heart Rate Monitoring". In: *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. ISSN: 2641-3604. Sept. 2022, pp. 1–4. DOI: 10.1109/BHI56158.2022.9926775.

[Hoc97] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

[Hol74] Klaus Holldack and Dieter Wolf. *Atlas und kurzgefasstes Lehrbuch der Phonokardiographie und verwandter Untersuchungsmethoden*. Thieme, 1974.

[Iwa21] Yuki Iwata, Han Trong Thanh, Guanghao Sun, and Koichiro Ishibashi. "High Accuracy Heartbeat Detection from CW-Doppler Radar Using Singular Value Decomposition and Matched Filter". In: *Sensors (Basel, Switzerland)* 21.11 (May 2021), p. 3588. ISSN: 1424-8220. DOI: 10.3390/s21113588. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8196719/ (visited on 08/21/2023).

[JC92] Lin JC. "Microwave sensing of physiological movement and volume change: a review". In: *Bioelectromagnetics* 13.6 (1992), pp. 557–565. DOI: "10.1002/bem.2250130610".

[Keb20] Mamady Kebe, Rida Gadhafi, Baker Mohammad, Mihai Sanduleanu, Hani Saleh, and Mahmoud Al-Qutayri. "Human Vital Signs Detection Methods and Potential Using Radars: A Review". In: *Sensors* 20.5 (2020). ISSN: 1424-8220. DOI: 10.3390/s20051454. URL: https://www.mdpi.com/1424-8220/20/5/1454.

[Kir93] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H. Hellhammer. "The 'Trier Social Stress Test' – A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting". In: *Neuropsychobiology*. Vol. 28. Issue: 1-2 ISSN: 0302282X. 1993, pp. 76–81. ISBN: 978-0-87421-656-1. DOI: 10.1159/000119004.

[Koe16] Alexander Koelpin, Fabian Lurz, Sarah Linz, Sebastian Mann, Christoph Will, and Stefan Lindner. "Six-Port Based Interferometry for Precise Radar and Sensing Applications". In: *Sensors* 16.10 (2016). ISSN: 1424-8220. DOI: 10.3390/s16101556. URL: https://www.mdpi.com/1424-8220/16/10/1556.

[Küd23] Arne Küderle, Robert Richer, Raul C. Sîmpetru, and Bjoern M. Eskofier. "tpcp: Tiny Pipelines for Complex Problems - A set of framework independent helpers for algorithms development and evaluation". In: *Journal of Open Source Software* 8.82 (2023), p. 4953. DOI: 10.21105/joss.04953.

[Li13] Changzhi Li, Victor M. Lubecke, Olga Boric-Lubecke, and Jenshan Lin. "A Review on Recent Advances in Doppler Radar Sensors for Noncontact Healthcare Monitoring". In: *IEEE Transactions on Microwave Theory and Techniques* 61.5 (May 2013). Conference Name: IEEE Transactions on Microwave Theory and Techniques, pp. 2046–2060. ISSN: 1557-9670. DOI: 10.1109/TMTT.2013.2256924. URL: https://ieeexplore.ieee.org/document/6504804 (visited on 09/27/2023).

[Mak21] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. "NeuroKit2: A Python toolbox for neurophysiological signal processing". eng. In: *Behavior Research Methods* 53.4 (Aug. 2021), pp. 1689–1696. ISSN: 1554-3528. DOI: 10.3758/s13428-020-01516-y.

[Mat00] G Matthews, B Sudduth, and M Burrow. "A non-contact vital signs monitor". en. In: *Crit. Rev. Biomed. Eng.* 28.1-2 (2000), pp. 173–178.

[Ord18] Celestino Ordóñez, Carlos Cabo, Agustín Menéndez, and Antonio Bello. "Detection of human vital signs in hazardous environments by means of video magnification". In: *PLoS ONE* 13.4 (Apr. 2018), e0195290. ISSN: 1932-6203. DOI: 10.1371/journal.

pone.0195290. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5895016/ (visited on 09/26/2023).

[Ric21] Robert Richer, Arne Küderle, Martin Ullrich, Nicolas Rohleder, and Bjoern M. Eskofier. "BioPsyKit: A Python package for the analysis of biopsychological data". en. In: *Journal of Open Source Software* 6.66 (Oct. 2021), p. 3702. issn: 2475-9066. doi: 10.21105/joss.03702. url: https://joss.theoj.org/papers/10.21105/joss.03702 (visited on 09/19/2023).

[Sam15] Michael Sampson and Anthony McGrath. "Understanding the ECG Part 2: ECG basics". In: *British Journal of Cardiac Nursing* 10.12 (Dec. 2015). Publisher: Mark Allen Group, pp. 588–594. doi: 10.12968/bjca.2015.10.12.588. url: https://www.magonlinelibrary.com/doi/abs/10.12968/bjca.2015.10.12.588 (visited on 09/18/2023).

[Sat23] Yasar Sattar and Lovely Chhabra. "Electrocardiogram". eng. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2023. url: http://www.ncbi.nlm.nih.gov/books/NBK549803/ (visited on 09/18/2023).

[Sch10] S. E. Schmidt, C. Holst-Hansen, C. Graff, E. Toft, and J. J. Struijk. "Segmentation of heart sound recordings by a duration-dependent hidden Markov model". eng. In: *Physiological Measurement* 31.4 (Apr. 2010), pp. 513–529. issn: 1361-6579. doi: 10.1088/0967-3334/31/4/004.

[Sch20] Sven Schellenberger, Kilin Shi, Tobias Steigleder, Anke Malessa, Fabian Michler, Laura Hameyer, Nina Neumann, Fabian Lurz, Robert Weigel, Christoph Ostgathe, and Alexander Koelpin. "A dataset of clinically recorded radar vital signs with synchronised reference sensor signals". en. In: *Scientific Data* 7.1 (Sept. 2020). Number: 1 Publisher: Nature Publishing Group, p. 291. issn: 2052-4463. doi: 10.1038/s41597-020-00629-5. url: https://www.nature.com/articles/s41597-020-00629-5 (visited on 08/30/2023).

[Shi19] Kilin Shi, Robert Weigel, Alexander Koelpin, Sven Schellenberger, Leon Weber, Jan Philipp Wiedemann, Fabian Michler, Tobias Steigleder, Anke Malessa, Fabian Lurz, and Christoph Ostgathe. "Segmentation of Radar-Recorded Heart Sound Signals Using Bidirectional LSTM Networks". In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. ISSN: 1557170X. IEEE, July 2019, pp. 6677–6680. isbn: 978-1-5386-1311-5. doi: 10.1109/EMBC.2019.8857863. url: https://ieeexplore.ieee.org/document/8857863/.

[Shi20] Kilin Shi, Sven Schellenberger, Christoph Will, Tobias Steigleder, Fabian Michler, Jonas Fuchs, Robert Weigel, Christoph Ostgathe, and Alexander Koelpin. “A dataset of radar-recorded heart sounds and vital signs including synchronised reference sensor signals”. en. In: *Scientific Data* 7.1 (Feb. 2020). Number: 1 Publisher: Nature Publishing Group, p. 50. ISSN: 2052-4463. DOI: 10.1038/s41597-020-0390-1. URL: https://www.nature.com/articles/s41597-020-0390-1 (visited on 08/30/2023).

[Shi21] Kilin Shi, Tobias Steigleder, Sven Schellenberger, Fabian Michler, Anke Malessa, Fabian Lurz, Nicolas Rohleder, Christoph Ostgathe, Robert Weigel, and Alexander Koelpin. “Contactless analysis of heart rate variability during cold pressor test using radar interferometry and bidirectional LSTM networks”. en. In: *Scientific Reports* 11.1 (Feb. 2021). Number: 1 Publisher: Nature Publishing Group, p. 3025. ISSN: 2045-2322. DOI: 10.1038/s41598-021-81101-1. URL: https://www.nature.com/articles/s41598-021-81101-1 (visited on 09/27/2023).

[Smi02] Julius O. Smith. *Digital Audio Resampling Home Page*. 2002.

[Spr16] David B. Springer, Lionel Tarassenko, and Gari D. Clifford. “Logistic Regression-HSMM-Based Heart Sound Segmentation”. In: *IEEE Transactions on Biomedical Engineering* 63.4 (Apr. 2016). Conference Name: IEEE Transactions on Biomedical Engineering, pp. 822–832. ISSN: 1558-2531. DOI: 10.1109/TBME.2015.2475278.

[Sri14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[Stü23] Sebastian Stühler. “Investigation of the Pre-Ejection Period as a Marker for Sympathetic Activity during Acute Psychosocial Stress”. bachelor’s thesis. Friedrich-Alexander Universität Erlangen-Nürnberg, 2023.

[Tu16] Jianxuan Tu and Jenshan Lin. “Fast Acquisition of Heart Rate in Noncontact Vital Sign Radar Measurement Using Time-Window-Variation Technique”. In: *IEEE Transactions on Instrumentation and Measurement* 65.1 (2016), pp. 112–122. DOI: 10.1109/TIM.2015.2479103.

[Vir20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[Wen21] Xuju Wang Wenjin Wang. *Contactless Vital Signs Monitoring*. Academic Press, 2021.

[Wie13] Uta S. Wiemers, Daniela Schoofs, and Oliver T. Wolf. "A friendly version of the Trier Social Stress Test does not activate the HPA axis in healthy men and women". en. In: *Stress* 16.2 (Mar. 2013), pp. 254–260. ISSN: 1025-3890, 1607-8888. DOI: 10.3109/10253890.2012.714427. URL: 10.3109/10253890.2012.714427 (visited on 10/17/2020).

[Wil17] Christoph Will, Kilin Shi, Robert Weigel, and Alexander Koelpin. "Advanced template matching algorithm for instantaneous heartbeat detection using continuous wave radar systems". In: *2017 First IEEE MTT-S International Microwave Bio Conference (IMBIOC)*. May 2017, pp. 1–4. DOI: 10.1109/IMBIOC.2017.7965797.

[Wil18] Christoph Will, Kilin Shi, Sven Schellenberger, Tobias Steigleder, Fabian Michler, Jonas Fuchs, Robert Weigel, Christoph Ostgathe, and Alexander Koelpin. "Radar-Based Heart Sound Detection". In: *Scientific Reports* 8.1 (July 2018), p. 11551.

[Xio17] Yuyong Xiong, Shiqian Chen, Xingjian Dong, Zhike Peng, and Wenming Zhang. "Accurate Measurement in Doppler Radar Vital Sign Detection Based on Parameterized Demodulation". In: *IEEE Transactions on Microwave Theory and Techniques* 65.11 (2017), pp. 4483–4492. DOI: 10.1109/TMTT.2017.2684138.

[Yoo21] Sungwon Yoo, Shahzad Ahmed, Sun Kang, Duhyun Hwang, Jungjun Lee, Jungduck Son, and Sung Ho Cho. "Radar Recorded Child Vital Sign Public Dataset and Deep Learning-Based Age Group Classification Framework for Vehicular Application". In: *Sensors (Basel, Switzerland)* 21.7 (Mar. 2021), p. 2412. ISSN: 1424-8220. DOI: 10.3390/s21072412. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8036835/ (visited on 08/30/2023).

Appendix A

Acronyms

DNN deep neural network

MAE mean absolute error

GUI graphical user interface

HPC high performance cluster

FAU Friedrich-Alexander Universität Erlangen-Nürnberg

PEP pre-ejection-period

CW continous-wave

ML medial-lateral

PA posterior-anterior

AE absolute error

HR heart rate

RLBM random large body movements

FMCW frequency-modulated continuous wave

ANS autonomic nervous system

biLSTM bidirectional long short-term memory

LSTM long short-term memory

HSMM hidden semi-Markov model

IBI interbeat interval

bpm beats per minute

FHS fundamental heart sound

bpm beats per minute

PCG phonocardiogram

ECG electrocardiogram

GPU graphics processing unit

ICG impedance cardiogram

HRV heart rate variability

RNN recurrent neural network

RMSE root mean squared error

TSST Trier social stress test

f-TSST friendly Trier social stress test

HoEnv homomorphic envelogram

HiEnv Hilbert envelope

PSDEnv power spectral density envelope

PUT person under test