University of Innsbruck
Faculty of Mathematics, Physics
and Computer Science

Department of Computer Science

Master Thesis

to obtain the academic degree

Master of Science

A Digital Talk Test for Assessing Exercise Intensity of Patients with Cardiovascular Diseases

by

Laura Geiger, BSc
(Matr.-Nr.: 11831841)

Submission Date:	February 29, 2024
Supervisor (UIBK):	Ass.-Prof. Dipl.-Ing. Clemens Sauerwein, PhD
Co-Supervisor (LBI DHP):	Mag. Dr. Daniela Wurhofer, Bakk. techn.
	Dr. Devender Kumar
	Dr.-Ing. Jan David Smeddinck, BSc, MSc

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Alle Stellen, die wörtlich oder inhaltlich den angegebenen Quellen entnommen wurden, sind als solche kenntlich gemacht. Die vorliegende Arbeit wurde bisher in gleicher oder ähnlicher Form noch nicht als Magister-/Master-/Diplomarbeit/Dissertation eingereicht.

Datum: ________________________________

Unterschrift: ________________________________

Contents

List of Figures

Abstract

Addressing cardiovascular diseases as a significant global health concern requires innovative approaches to enhance patient care and rehabilitation. One of these approaches is the development of a digital version of the Talk Test which enables patients to self-assess their intensity through speech during workouts. Furthermore, the automated assessment of exercise intensity through speech related to the Talk Test offers a compelling and engaging research area. Currently, no publicly available machine learning approaches address the automated exercise intensity estimation. In this thesis a prototype of the Digital Talk Test, called ***aktivtalk***, is created and used in a comparative study to collect self-assessed voice samples. The usability of the ***aktivtalk*** application achieves a higher user satisfaction compared to the non digital approach in the study. Furthermore, an initial machine learning model is created to predict the exercise intensity zone without the need for self-assessment. The performance of the model with the self-assessed labels even slightly outperformed a traditional age and pulse based labeling approach. In summary, this thesis emphasizes the potential of the Digital Talk Test as a valuable tool for assessing and managing exercise intensity in patients with cardiovascular diseases, highlighting its reliability, usability, and potential for future development in digital healthcare.

1. Introduction

Low physical fitness levels increase the risk of health issues, including cardiovascular diseases (CVDs). According to the World Health Organization, CVDs account for an estimated 17.9 million annual deaths, making them the leading global cause of death [13, 12]. However, around 14% of the total mortality risk can be prevented by working out regularly [2, 5]. Physical exercise is also highly recommended for patient rehabilitation, however, measuring only heart rate is insufficient, not only for patients taking beta-blockers but also for high-level athletes [19]. Bias could be caused by beta-blockers because they reduce the heart rate and cardiac stress during exercise. Therefore, exercise tests that are not based on the heart rate, such as the Talk Test, are required [7, 28]. The Talk Test is a non-invasive procedure for self-assessing the exercise intensity area during workouts. After reading out loud a short paragraph the patient rates their comfort of speech by answering with ”Yes”, ”Not sure” or ”No” to the question if talking was still comfortable. This master thesis is done in cooperation with the Ludwig Boltzmann Institute for Digital Health and Prevention (LBI DHP), which focuses on the prevention and rehabilitation of cardiovascular diseases through physical activities. The research questions guiding this master thesis are as follows:

RQ1: What is the Talk Test, and is it a reliable tool for assessing exercise intensity in patient rehabilitation?

RQ2: How can the Talk Test be transformed into a digital format suitable for patients diagnosed with cardiovascular diseases?

RQ3: Does the digital version of the Talk Test have a high level of satisfaction when considering the system's usefulness, information quality, interface quality, and trust from the user's perspective?

RQ4: Can an initial machine learning model effectively detect and classify exercise intensity levels from audio files collected by a digital version of the Talk Test, and if so, what is the preliminary performance of such a system?

The scope of this master thesis includes the development of a prototype of a Digital Talk Test since there currently exists no such tool on the market of digital health technologies. This innovative approach is used in a comparative study to collect voice samples during different exercise intensity stages from participants. The exercise intensity is self-assessed by the participants and labels the recorded voice samples. These samples are collected in a publicly available dataset. Furthermore, the study includes user testing in order to investigate the usability of the application. The data is evaluated and used to train an initial machine learning model that predicts exercise intensity from recorded voice samples. The goal is to establish a foundation and initial approach for a digital version of the Talk Test with self assessment and machine learning based intensity assessment.

The thesis is structured as follows. Chapter 2 delves into the background, exploring the Talk Test, addressing RQ1 and identifying research gaps. In Chapter 3, the research methodology is outlined, employing the design science research method for prototype development addressing RQ2, conducting a literature review, explaining the user study and machine learning-based intensity classification methodologies. Chapter 4 presents research outcomes, discussing findings from the systematic literature review on the Talk Test and describing the ***aktivtalk*** prototype. Chapter 5 evaluates research outcomes, presenting user study findings, analyzing the synchronization data and assessing machine learning-based intensity classification by addressing RQ3 and RQ4.

2. Background

2.1. Talk Test

Valuable insights into the Talk Test and its reliability have been gained through an exploration of its previous use in patient rehabilitation. For addressing RQ1 previous literature on this topic is essential. The background not only shows key developments in the application of the Talk Test but also establishes a robust foundation for the upcoming chapters.

2.1.1. Procedure

In the health domain, the concept underlying the Talk Test (TT) traces its roots back to approximately 1937 when Scottish mountaineers were advised by John Grayson to ”climb no faster than you can talk”. This principle is now applied to assess exercise intensity in patients. To determine if talking is still possible, the patient reads a standard paragraph (about 30-50 words) and then answers a question about whether speaking was comfortable or not. If the patient answers with ”yes” then the test continues and the same procedure is repeated after some time. If the patient answers with ”yes, but” or ”I'm not sure”, the stage is marked as equivocal. The test usually finishes if the answer is ”no” [3]. The different stages of the Talk Test help to assess the exercise intensity without using the pulse as an intensity measure. Although the Talk Test does not fully align with subjective intensity measurements like the Ventilatory Threshold, it is a reliable tool for patient rehabilitation.

The typical procedure of the Talk Test is illustrated in Figure 2.1 below:

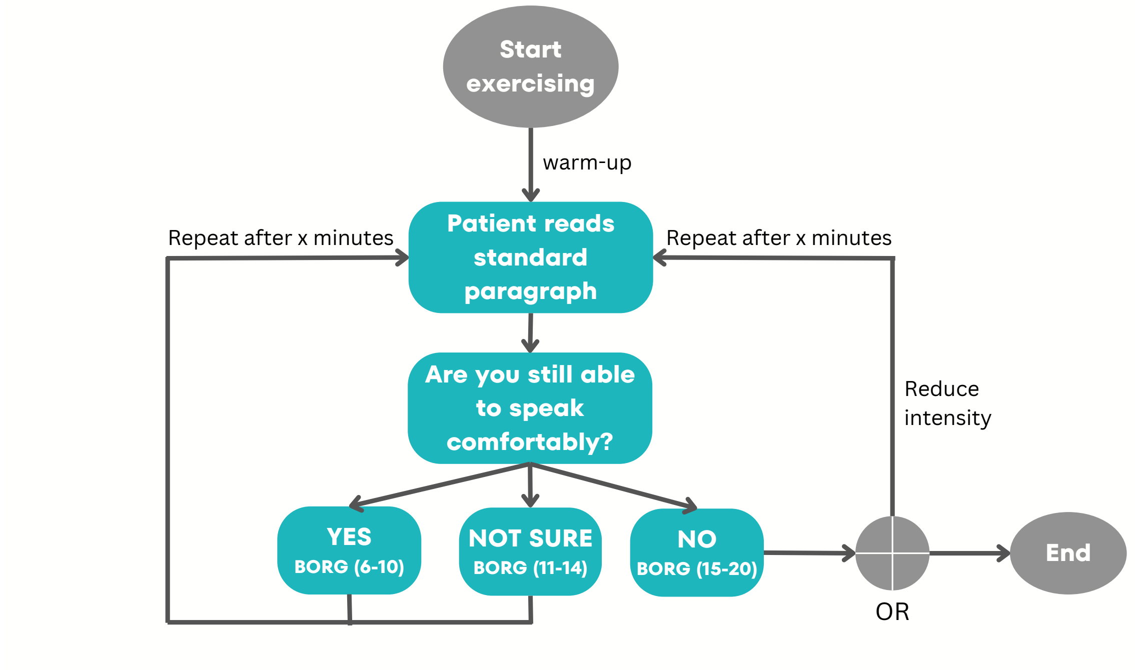


Figure 2.1.: Talk Test procedure

In the conducted studies of the literature review, patients are using a cycle-ergometer, beginning with a warm-up at 0 watts for a few minutes. Consequently, patients are required to read a standard 30-word paragraph aloud. The question if the patient is still able to speak comfortably or not is then asked. They can answer with "Yes" (positive stage) or "Not sure" (equivocal stage) indicating light to medium intensity, or "No" (negative stage) indicating high intensity. If the negative stage is reached, the workout is terminated. In the context of this thesis the assessment with "Yes", "No" and "Not sure" is indicated as **YNNS**.

According to Bok et al. [3], there is a mapping between the initial stages of the Talk Test and the rating of perceived exertion scale, as illustrated in Figure 2.2. For this master thesis the mapping was expanded by adding the light, medium and high intensity zone according to the areas divided by the thresholds. From the mapping it can be derived that the zones of the Talk Test can be assessed with a more detailed partitioning.

- Positive stage (light intensity): ”yes”, ”6”, ”7”, ”8”, ”9”, ”10”
- Equivocal stage (medium intensity): ”not sure”, ”11”, ”12”, ”13”, ”14”
- Negative stage (high intensity): ”no”, ”15”, ”16”, ”17”, ”18”, ”19”, ”20”

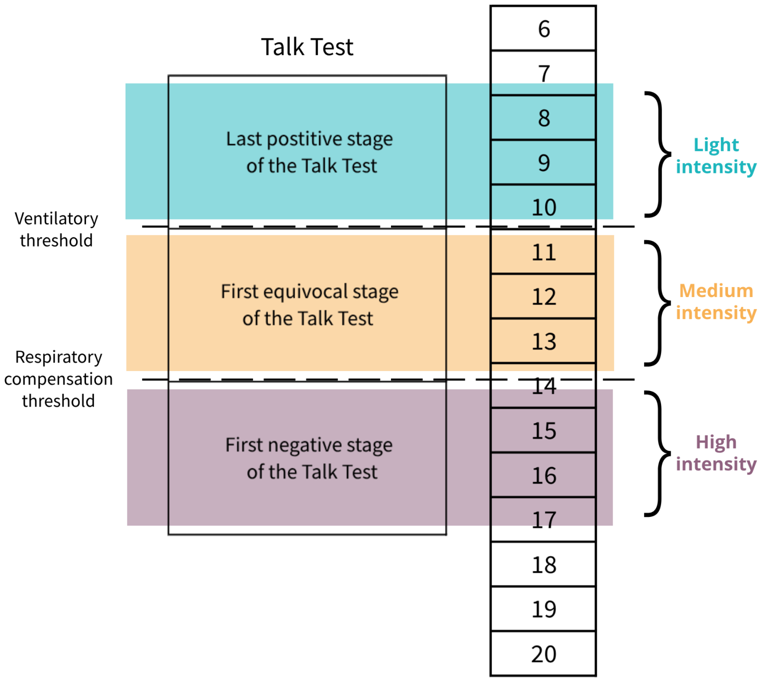

Talk Test
Last postitive stage of the Talk Test
First equivocal stage of the Talk Test
First negative stage of the Talk Test
Ventilatory threshold
Respiratory compensation threshold
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Light intensity
Medium intensity
High intensity

Figure 2.2.: Talk Test and BORG Scale mapping
(Own illustration based on Bok et al. [3])

2.1.2. Ventilatory Threshold

Exercise intensity can be measured by the physiologic and metabolic response like the ventilatory threshold (VT). The ventilatory threshold can be divided into the first and second ventilatory threshold. The first ventilatory threshold (VT1) is the point where the exercise intensity and the breathing rate increase. When the VT1 is reached then talking comfortably is no longer possible. The second ventilatory threshold (VT2) indicates the point when the person heavily breathes and is no longer able to speak [27].

2.1.3. State of the Art

The most recent trial about the TT by Orizola-Cáceres et al. [16] compares the utility of the traditional talk test (TTT) and alternative talk test (ATT) for the determination of aerobic training zones in overweight and obese patients. In zone 1 the intensity is smaller than the VT1 and for zone 2 the intensity is between VT1 and VT2. In zone 3 the intensity is higher than VT2. 19 obese/overweight subjects aged 34.9 ± 6.7 years were included in the trial. The participants underwent an incremental ergometric test for maximal oxygen consumption on a cycle. At the end of each stage, the participants had to answer the question "Was talking comfortably?" with "Yes", "No" or "I don't" for the TTT after reading out loud a text of 40 words. Alternatively, the participants underwent the ATT where they had to identify the comfort of talking through a 1 to 10 numeric perception scale, which is a visual analog scale (VAS). The TT consists of a 10-minute warm-up, followed by load increments every 3 minutes. A relation between the power output of the first no (FN) and the power output at the VT2 was found in the TTT. Furthermore, there was an agreement between the power output of the last yes (LY) of the TTT and the power output at the VAS 4-5 (talking was slightly difficult) of the ATT. However, there was no relation between the power output of the LY of the TTT and the power output at the VT1. Between the power output at the VAS 2-3 of the ATT and the power output of the VT1 there was an agreement. However, the TTT failed to find an agreement between the different answers and the power output at VT1. The results of this study indicate that the ATT is a more valid and better-tolerated measure of exercise intensity than the traditional talk test. The paper concludes that the ATT is a low-cost and easy-to-apply tool for measuring the exercise intensity for exercise prescription in overweight/obese patients. However, the TTT could under- or overestimate the physical effort [16].

The reliability of the graded cycling test in comparison with the 30-s chair-stand test in men with prostate cancer on androgen deprivation therapy was measured by Aabo et al. [1]. 60 men aged 70.8 ± 6.6 had to perform a GCT-TT and 30s-CST twice. The GCT-TT yielded a relative reliability of 0.90 with a confidence interval of 95% (0.94-0.98). The GCT-TT started with a low intensity of 15 watts and 60 rounds per minute for 2 minutes. The intensity was increased by 15 watts every minute. The men had to recite a text consisting of 30 words during the last 10 seconds of every minute and afterward, they had to answer the question "Are you still able to speak comfortably?" with the following responses: "yes", "unsure" or "no". After the GCT-TT the 30s-CST was executed and both tests were repeated with identical conditions on the same day. There was a high correlation between the output of the first and second tests of the GCT-TT ($r = 0.901$). Furthermore, a high $ICC_{2.1}$ value of 0.9 (95% CI 0.84-0.94) was recorded for the GCT-TT. In conclusion, this study demonstrated that the graded cycling test with TT and 30-s chair-stand test is a reliable method of assessing exercise capacity in men with prostate cancer on androgen deprivation therapy in the clinical setting [1].

Another study by Sørensen et al. [23] addressed the validity of the TT as a method to estimate ventilatory threshold and guide exercise intensity in cardiac patients. 20 cardiac patients with a mean age of 65 ± 8.5 years performed two exercise tests with an identical ramp protocol on a cycle ergometer on the same day. The first one was a submaximal effort test on a cycle to assess exercise intensity using the TT. The participants had to read out loud a text of 30 words every minute within 10 seconds. After that, the typical TT question procedure was done and the following responses were acquired: TT_{pos} positive response, TT_{eq} equivocal response and TT_{neg} negative response. The second test identified the VT with a cardiopulmonary exercise test using breath-by-breath gas analysis. When evaluating the Vo_w and workload with a statistical analysis, the intensity at TT_{eq} and TT_{neg} was not significantly different from the VT. However, the intensity at TT_{pos} was below the VT and the limits of agreement demonstrated wide ranges. Nonetheless, the stages of the TT were in accordance with the prescribed exercises for the patients and the intensity at TT_{pos} and TT_{eq} indicate moderate-intensity continuous training. Therefore, the TT can be useful for prescribing exercise intensity for cardiac patients [23].

Krawcyk et al. [11] concluded that the graded cycling test combined with the TT is a reliable tool for monitoring cardiovascular fitness in patients with minor strokes. A pilot study was done before the main study in order to practice the test procedure. The main study included 60 participants with a mean age of 67 and a range of 44-85. The cycling test was performed two times on the same day in order to test the reliability. At the beginning of each test the participants cycled at a low intensity of 15 watts and 60 rounds per minute for two minutes. After that, the intensity was increased every minute by 15 watts and in the last 10 seconds, the participants had to read out loud a text of 30 words. The TT was done and the three responses were identified: Test+ (”yes”), Test ± (”unsure”) and Test- (”no). The reliability of the TT was measured by comparing the outcome of the first and second test results. The $ICC_{2.1}$ was around 0.92 (95% CI 0.81-0.95) for TT+ and around 0.97 (95% CI for 0.95-0.98) for TT-. Furthermore, the patients cooperated well with TT and the test protocol. Because of its high reliability, good acceptance, and user-friendliness, the test is suitable for research and clinical practice [11].

Furthermore, Nielsen and Vinther [15] conducted a study about the responsiveness of the graded cycling test combined with the TT (GCT-TT) in cardiac rehabilitation. The changes of the GCT-TT were measured after an 8-week cardiac rehabilitation program. 93 patients aged 63.3 ± 9.7 years had to perform a GCT-TT on a stationary ergometer cycle before entering the rehabilitation. The same test was executed at the last or second to last training session. During the rehabilitation, the patients participated in an 8-week training program with 2 weekly 1.5-hour training sessions. Moreover, the patients were encouraged to do home exercise sessions like walking or cycling. Some patients were excluded from the second test because of illness and outliers were removed. Therefore, 85 patients were taken into consideration for the results. The increase from the pretest to the posttest was about 18.1 ± 21.1 W.

Furthermore, the patients had to answer a question about their level of physical fitness at the end of the rehabilitation, compared to when they started it. 17 patients said that there was no to minor change. 26 patients stated that there was some change and for 37 there was a major change. This could also be observed when dividing the patients into three subgroups regarding the different answers. There was a significant increase in the change mean for the patients that observed some change and major change. Therefore, a strong agreement could be found between the test changes and the changes in the level of fitness perceived by the patients [15].

Nielsen at el. [14] investigated the reliability of the graded cycling test combined with the TT for patients with ischemic heart disease. 64 patients aged 36 to 82 were included in the study. The patients were introduced to the test before the actual testing and they were told that the purpose of testing was to investigate the test and not the fitness level. This was done in order to reduce bias. Two GCT-TTs were performed on a cycle ergometer by the patients on the same day with 2 hours between the tests. The GCT started with a warm-up of 2 minutes at the lowest intensity. The intensity was increased by 15 watts every minute and the patients had to recite a 30-words text during the last 10 seconds of every minute. After that, the typical TT procedure was done and the patients had to give a positive (TT+), negative (TT-) or equivocal (TT±) answer to the question of whether they are still able to speak comfortably or not. The data of the workload (W) was collected for each answer. Furthermore, two experienced physiotherapists rated the point in time at which the patients could no longer speak comfortably in a pilot study that included 8 healthcare professionals and 4 patients with ischemic heart disease. The ratings of the physiotherapists were not seen by the patients in order to minimize bias. For the results of the TT and the ratings of the physiotherapists, relatively high ICC values were observed. Moreover, the correlation between all the TT answers was excellent. The GCT-TT was well adapted by the patients and there was only a small measurement error. Furthermore, Nielsen et al. concluded that the GCT combined with the TT is a safe and efficient tool for exercise prescription. However, the role of the TT during unsupervised physical activities in patients with cardiac diseases remains to be investigated [14].

2.2. Research Gap

The existing literature and Chapter 4.1 highlight a research gap in the field of exercise intensity assessment, particularly in cardiovascular health management. The Talk Test has been used for many decades and its reliability has been proven many times, however, there is currently no digital implementation of the Talk Test available. Furthermore, the automated estimation of exercise intensity through speech eliminates the self-assessment and is not well researched yet. A machine learning based approach is needed to solve this task. Since there is currently no publicly available dataset that contains audio files featuring patients' self-assessed intensity zones, it becomes necessary to collect data for the machine learning, which further compounds the research gap.

3. Research Methodology

3.1. Design Science Research

The Design Science Research (DSR) method serves as the overarching framework for this thesis. DSR is particularly suited for projects focused on creating innovative artifacts to address real-world problems. In the context of this research, the artifact is the ***aktivtalk*** application, designed to enable patients to self-assess exercise intensity through speech during workouts. The DSR process involves a cyclical iteration of design, development and evaluation, aligning with the iterative nature of creating and refining a digital health technology. This processes contributes to RQ2, since it enables the development of a digital version of the Talk Test based on the medical background in cardiovascular diseases.

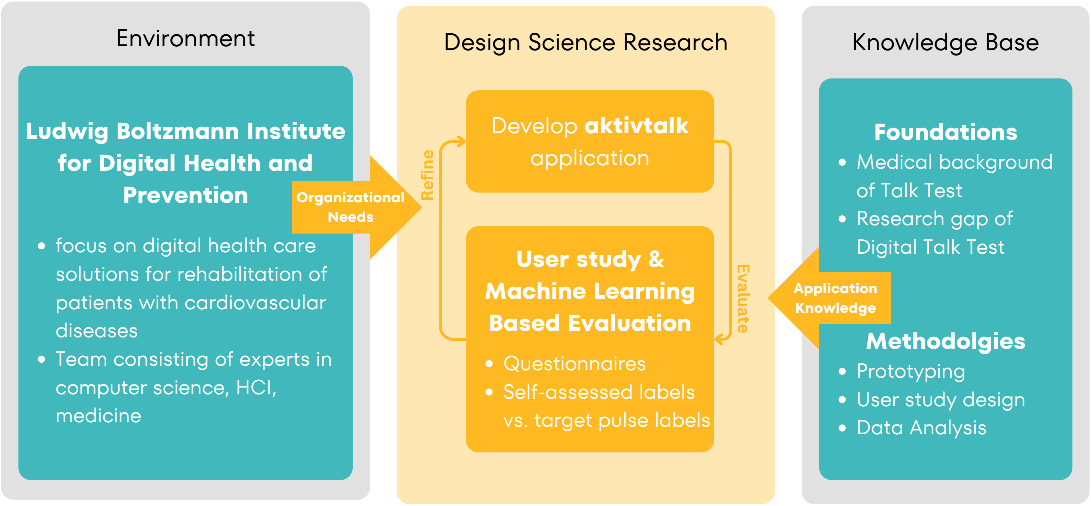

Environment
Ludwig Boltzmann Institute for Digital Health and Prevention
• focus on digital health care solutions for rehabilitation of patients with cardiovascular diseases
• Team consisting of experts in computer science, HCI, medicine
Organizational Needs
Design Science Research
Refine
Develop aktivtalk application
User study & Machine Learning Based Evaluation
• Questionnaires
• Self-assessed labels vs. target pulse labels
Evaluate
Application Knowledge
Knowledge Base
Foundations
• Medical background of Talk Test
• Research gap of Digital Talk Test
Methodolgies
• Prototyping
• User study design
• Data Analysis

Figure 3.1.: Design Science Research method

3.1.1. Environment

The research environment is the Ludwig Boltzmann Institute for Digital Health and Prevention (LBI DHP). The institute's commitment to cardiovascular diseases prevention and rehabilitation provides a rich setting for the development and evaluation of the ***aktivtalk*** application. Collaboration with LBI DHP ensures access to expertise in digital health and a relevant context for the deployment of the artifact.

3.1.2. Knowledge Base

The knowledge base for this research is derived from existing literature and a systematic literature review. The World Health Organization's alarming statistics on CVDs underscore the importance of innovative solutions. The literature review provides insight into the theoretical background of the Talk Test and shows its reliability.

3.1.3. Artifact Development

The central artifact of this research is the ***aktivtalk*** application, serving as a prototype for the Digital Talk Test. The development process involves creating a user-friendly interface and integrating self-assessment features and mechanisms throughout the workout. The design choices are influenced by the need for inclusivity, considering patients on beta-blockers.

3.1.4. Evaluation

The usability of the ***aktivtalk*** application, including usefulness, information quality, interface quality and trust, is evaluated in a user study. Besides the data from questionnaires and interviews, voice samples and pulse recordings are collected. These voice samples and pulse recordings are synchronized in order to have a deeper insight into the data and compute the target heart rate zones for evaluating the machine learning model. The initial machine learning model is trained using the collected voice samples. The model's performance is assessed, comparing its predictions with both self-assessed labels and a traditional age and pulse-based labeling approach. This evaluation provides insights into the feasibility and accuracy of the intensity zone estimation without self-assessment.

3.2. Systematic Literature Review

3.2.1. Search Strategy

The search was performed in Google Scholar and Springer based on the PICOC scheme in Table 3.1. PubMed is usually used for literature in medicine and life sciences, however, in an initial iteration, Google Scholar and Springer identified more relevant literature. Springer, was selected to access specialized content with a high standard of research quality, while Google Scholar offers a wide and diverse range of scholarly sources. By combining both databases, it was aimed to achieve a balance between depth and breadth, ensuring that the systematic literature review captured a comprehensive view of existing research on the Talk Test.

Population	Literature in Talk Tests for Patient Rehabilitation
Intervention	The Realization of the Talk Test in different case studies
Comparison	Case studies together with their evaluation
Outcome	Several Observations regarding reliability of different case studies, limitations and open challenges in this field
Context	State of the Art of Talk Tests for Patient Rehabilition

Table 3.1.: PICOC scheme

The search term "Talk Test" was chosen to be broad on purpose. Using a wide term was important to make sure to include any relevant literature on the Talk Test in patient rehabilitation.

Furthermore, for the Watson and Webster approach was combined with the Kitchenham approach. First, the starting papers were identified. After that, the citations of the starting papers were reviewed (going backward) and articles that cite the starting papers were identified (going forward).

3.2.2. Selection Criteria

Only papers in the English language created after 2017 that had the term "Talk Test" in their title were included in the first iteration. Moreover, papers about the healthy population were excluded. As mentioned in the introduction the TT is especially useful for patients with CVDs. However, this systematic literature review also includes patients with other diagnoses, in order to get a broader application area of the TT.

The selection procedure was planned as follows:

1. Identify papers with search term and full-text search

2. Screen titles

3. Identify quality in SCImago journal rank

4. Screen content

3.2.3. Quality Assessment

The quality of the literature was identified by using the SCImago journal rank [10]. The rank is built on the number of citations received by a journal and the prestige of the journals where the citations come from. For this Systematic Literature Review, papers from the first and second quartiles were included and from the third and fourth quartiles were excluded.

3.3. User Study for the Digital Talk Test

The procedure of the user study can be seen down below in Figure 3.2. Before the 30-minute workout a pre-questionnaire about demographic data, physical activity and smartphone use for health-related purposes is conducted. The workout is subdivided into three different methods which are permutated across the participants. After the workout, a post-questionnaire about the usability of the ***aktivtalk*** application is conducted. This is followed by a short interview.

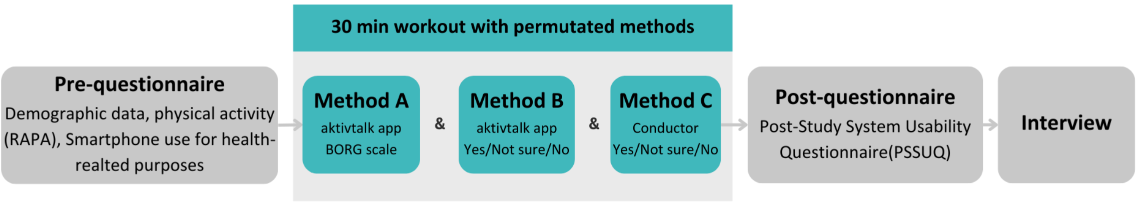

30 min workout with permutated methods
Pre-questionnaire
Demographic data, physical activity (RAPA), Smartphone use for health-realted purposes
Method A
aktivtalk app
BORG scale
&
Method B
aktivtalk app
Yes/Not sure/No
&
Method C
Conductor
Yes/Not sure/No
Post-questionnaire
Post-Study System Usability Questionnaire(PSSUQ)
Interview

Figure 3.2.: Study procedure

3.3.1. Participants and Sample Size

The study participants are healthy individuals who are able to use a smartphone and can engage in regular physical activities. They are recruited through the use of snowball sampling, primarily targeting individuals within the researcher's circle of acquaintances. Efforts are made to ensure a balanced representation of participants based on their self-identified gender. This approach aims to reflect the diversity of the group in an equitable manner. The data collection phase involves a total of 20 participants.

3.3.2. Location

The data collection phase occurs in a controlled and quiet indoor environment. To ensure the absence of any external noise disturbances, it is essential that only the conductor and the participant are present in the room. The facility must be equipped with an ergometer or similar indoor cycling training equipment and have windows. No additional specifications or prerequisites are required for the study.

3.3.3. Informed Consent

The workout session and data collection solely proceeds upon the participant's explicit consent, as outlined in Appendix D. Informed consent is provided in the German language, considering that the native language of each participant is German.

3.3.4. Methods

For the study, three methods are distinguished in order to compare the ease of use, trust and preference against each other. The methods are displayed below in Table 3.2:

	Method A	Method B	Method C
Mobile Application	BORG Scale	Yes/Not sure/No	
Conductor guided			Yes/Not sure/No

Table 3.2.: Study methods

Each participant trains for 10 minutes with each method. In total, the workout lasts 30-35 minutes. The methods are permutated across the participant in order to mitigate any bias by a possible learning effect or the increasing intensity of the workout. For the set of the three methods A,B,C there are the following six permutations: (A,B,C), (A,C,B), (B,A,C), (B,C,A), (C,A,B), (C,B,A). Since there are twenty participants, each permutation is used three times, except (A,B,C) and (A,C,B) are used four times.

3.3.5. Pre-Questionnaire

Down below the questionnaire for collecting information regarding the participants demographics, physical activities and smartphone use is shown.

Demographic data

1. What is your age?
2. What gender do you identify as?

Rapid Assessment of Physical Activity

The Rapid Assessment of Physical Activity (RAPA) [18] questionnaire is a tool designed to assess an individual's level of physical activity. It is particularly useful in healthcare settings to quickly and effectively measure a person's activity habits. The questionnaire consists of 9 questions that assess a person's physical activity level, including both aerobic and muscle-strengthening activities. It is designed to be self-administered and is relatively quick to complete. The highest score is used to classify an individual into different activity levels, which can be categorized as ”active”, ”under-active regular”, ”under-active regular - light activities”, ”under-active” and ”sedentary” [18]. The questionnaire can be seen in Appendix A.

Smartphone use (health apps)

1. Do you use your smartphone for health-related purposes, such as using digital health apps, tools, or services (Steps counter, pulse tracker, etc.)?
2. If yes, what apps?

3.3.6. Equipment

In conducting this study, the requisite equipment plays an important role in ensuring precise data collection and reliable results. The selected instrumentation serves as the foundation for capturing essential variables, fostering the accuracy and validity of our research outcomes:

- Smartphone 1
- Smartphone 2
- Polar OH1+ optical heart rate sensor
- Lavaliere microphone
- Smartphone holder for bike handlebar
- Smartphone tripod

3.3.7. Preparation

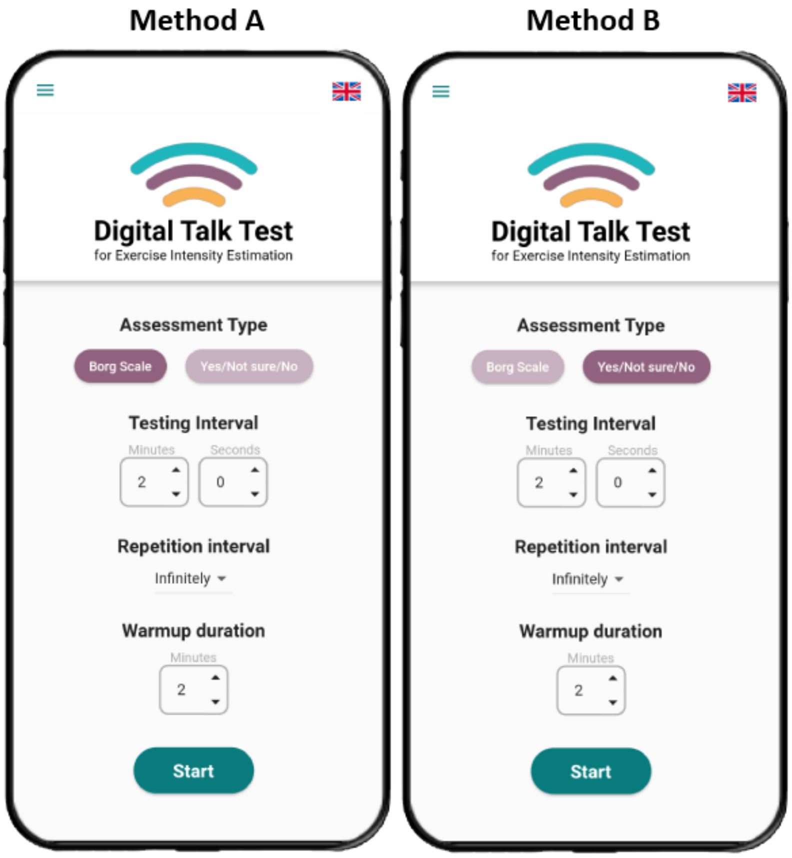
In preparation for the study, attention is given to ensure optimal conditions for participants. The steps can be seen down below:

1. Spin bike is correctly adjusted according to the body size of the participant

2. Smartphone 1 is attached to the handlebar with holder

3. Smartphone 1 settings:

   aktivtalk is active



Method A

Assessment Type: Borg Scale

Testing Interval: Minutes 2, Seconds 0

Repetition interval: Infinitely

Warmup duration: Minutes 2

Method B

Assessment Type: Yes/Not sure/No

Testing Interval: Minutes 2, Seconds 0

Repetition interval: Infinitely

Warmup duration: Minutes 2

Figure 3.3.: ***aktivtalk*** settings

4. Smartphone 1 screen timeout is turned off or set to high amount of minutes

5. Smartphone 2 is positioned on a tripod with a maximum distance of 1,5 meters next to the participant

6. Polar OH1+ is attached to left arm, connected to mobile application and turned on

7. No resistance is set on spin bike

8. Sufficient water supply is ensured for participant
9. Participant wears lavaliere microphone (plugged into smartphone with ***aktivtalk*** app)
10. Participant is ready to start the workout

3.3.8. General Guidelines

When conducting the study it is important to adhere to the following guidlines:

- The participant can choose the intensity, however, there should be at least 5 recordings of each intensity level.
- The workout should take 30 to 35 minutes including the warmup.
- The patient can cancel the workout at any time.

3.3.9. Opening

Below, the introduction and instructions of the study conductor are presented.

Interviewer:

Hello and thank you for participating in my study for collecting data for my master thesis. The master thesis and this study includes the testing of a digital version of the Talk Test for assessing exercise intensity.

(Depending on consent) A recording device will be used to record your voice during reading out loud. These samples will then be used for a machine learning model which automatically detects the exercise intensity area from the collected audio samples.

Your task is to train on the spin bike for up to 35 minutes and the intensity can be freely chosen by you. Please make sure that you do not only train in a very light area, but that you may also get ”out of breath” at times. The test is not intended to measure your maximum performance, but to record your voice in the different training ranges. If you do not feel well, you can stop the test at any time.

You will be using the following three methods for estimating your exercise intensity:

Method A:

You will be using the activtalk app. From the beginning of the test, you will be asked to read out a paragraph of about 30 words every two minutes. You will then be asked whether you can still speak comfortably. You can rate yourself on a scale of 6 - 20.

Method B:

You will be using the activtalk app. From the start of the test, you will be asked to read out a paragraph of about 30 words every two minutes. You will then be asked whether speaking is still comfortable for you. You can answer ”Yes”, ”No” or ”Not sure”.

Method C:

I will guide you through the workout. From the start of the test, you will be asked to read out a paragraph of about 30 words every two minutes. You will then be asked whether speaking is still comfortable for you. You can answer ”Yes”, ”No” or ”Not sure”.

Do you have any questions?

3.3.10. Workout Start

The start of the workout depends on the first method of the permutation. Therefore, it is important to distinguish between the steps to follow when initiating the workout.

Starting with Method A or B

- Conductor starts the voice recording on Smartphone 2
- The participant starts the workout by clicking on the “Start” button on Smartphone 1 and then says “Start” and claps once simultaneously.
- The conductor clicks on the start button in the polar beat application simultaneous to the clap.

Starting with Method C

- Conductor starts the voice recording on Smartphone 1 and 2
- Participant starts the workout with saying “Start” and claps once simultaneously.
- The conductor clicks on the start button in the polar beat application simultaneous to the clap.

3.3.11. Workout

During the workout, voice samples of the participants are gathered. The methods are changed every 10 minutes by the conductor due to the participants’ permutation order.

Method A and B

- The participant is guided through the workout by the application

Method C

1. 2-min warmup
2. Conductor asks the participant to read out loud a 30-words paragraph

3. The conductor asks the participant if they are still able to speak comfortably Method C: Participant can answer with yes, not sure or no

4. Depending on the answer the conductor gives the participant feedback

   - Answer "Yes": Light intensity → continue or increase watts
   - Answer "Not sure": Medium intensity → continue or increase watts
   - Answer "No": High intensity → continue or decrease watts if overexerted

5. Repeat steps 1 to 4 until the workout is over

3.3.12. Workout End

If the workout has already reached a duration of 30 to 35 minutes, and a sufficient amount of voice data has been collected, the session is terminated. Subsequently, the following steps are to be carried out by the conductor:

1. Sanitize earphones
2. Sanitize spin bike

3.3.13. Post-Questionnaire

The post-questionnaire is carried out after the participant ends the workout and is ready to answer questions about the perceived usability of the ***aktivtalk*** application.

Post-Study System Usability Questionnaire

The Post-Study System Usability Questionnaire (PSSUQ) is a widely recognized and standardized questionnaire used to assess the usability and user satisfaction of a system or product, particularly in the context of human-computer interaction (HCI) and user experience (UX) research. It is employed to gather feedback from users after they have interacted with a system or software application, enabling the evaluation of its overall usability and the users' perceptions of its effectiveness, efficiency, and satisfaction [17]. There are three sub-scales that provide a more detailed breakdown of different factors:

- System Usefulness (SYSUSE): average scores of questions 1 to 6
- Information Quality (INFOQUAL): average scores of questions 7 to 12
- Interface Quality (INTERQUAL): average scores of questions 13 to 15
- Overall: average scores of questions 1 to 16

The scoring range for PSSUQ begins with 1 (strongly agree) and ends with 7 (strongly disagree). Lower scores indicate better performance and satisfaction. However, it is important to note that 4 is considered neutral and may not necessarily represent the average score. Additionally, a score below 4 does not necessarily imply that your website, software, system, or product has performed above average. Moreover, the post-questionnaire includes open-ended questions that inquire about participants' preferences, dislikes, and suggestions for further enhancing the system. The questionnaire can be seen in Appendix B.

3.3.14. Pre- and Post-Questionnaire Remarks

Participants responding to the questionnaire may be susceptible to social desirability bias, as described by Grimm [8]. This bias relates to individuals' tendency to choose socially desirable or acceptable answers instead of providing genuine reflections of their thoughts and emotions. It is possible that participants might try to please the conductor, for example, in order to avoid negative judgments. To counteract this bias, the questionnaires are anonymous. Additionally, it is crucial to acknowledge the potential impact of the Hawthorne Effect when administering the questionnaires to the participants. The Hawthorne Effect refers to the phenomenon wherein participants modify their behavior and responses due to their awareness of being observed.

As participants are aware of being evaluated, their attitudes and responses may be influenced, potentially affecting the reliability of the survey findings as explained by Sedgwick [20]. To mitigate this effect, the conductor exits the room while the questionnaire is being completed.

3.3.15. Interview

The following questions are asked after the post-questionnaire is filled in and recorded with Smartphone 2:

- What do you think is the best aspect of this software, and why?
- What do you think needs most improvement, and why?
- Do you trust the yes, no or not sure assessment when using the ***aktivtalk*** application?
- Do you trust the BORG scale assessment when using the ***aktivtalk*** application?
- Do you trust the yes, no or not sure assessment when guided by the conductor?

3.3.16. Collected Data

The data to be collected in the study can be seen in Figure 3.4 below. The data of the participants includes the single audio samples from Smartphone 1, the continuous audio files (CAFs) from Smartphone 2 and the information of the pulse during the workout. Moreover, data from the questionnaires and interviews is extracted.

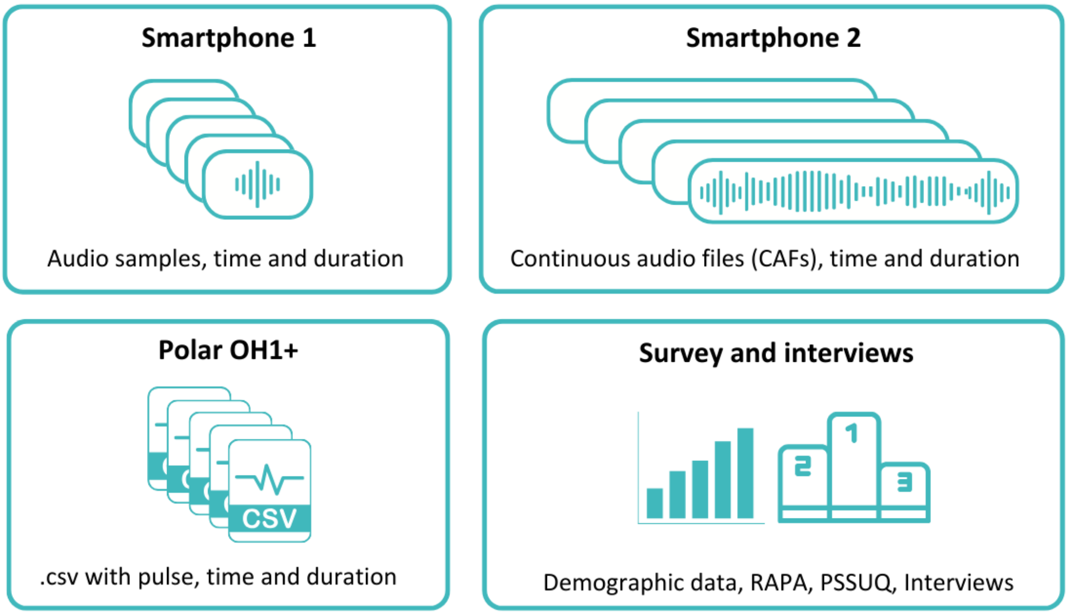

Smartphone 1
Audio samples, time and duration
Smartphone 2
Continuous audio files (CAFs), time and duration
Polar OH1+
CSV
.csv with pulse, time and duration
Survey and interviews
1
2
3
Demographic data, RAPA, PSSUQ, Interviews

Figure 3.4.: Overview collected data

3.3.17. Test Trial

A test trial prior to the study conduction is essential to validate the viability of the protocol and ensure the comprehensive collection of all the relevant data. By simulating the actual study conditions, including the use of the ***aktivtalk*** application, the recording of voice samples and the pulse during the workout, the test trial helps to identify any potential challenges or shortcomings in the methodology. This ensures that the actual study conduction is running smoothly and produces reliable results.

3.3.18. Data Storage and Management

In general, all information gathered during this project is treated with confidentiality and stored in compliance with GDPR regulations. Any researchers with access to the data must adhere to relevant national data protection laws and the GDPR. Furthermore, the LBI DHP in Salzburg, where the data is stored on external devices, is not accessible to the public. Only authorized employees are permitted entry. Moreover, the password-protected hardware is securely stored at the institute. Access is granted solely to the project lead and authorized researchers who are involved in project administration and data analysis. The hardware stores the following information in separate files, each of which is password-protected for added security:

- **Personal data:** Participants' personal data is collected in a questionnaire at the beginning of the study. Furthermore, demographic data (age, gender), physical activities and frequency, smartphone use (digital health apps) are collected.
- **Audio recordings** are captured during the exercise session while participants engage in the Talk Test. The collected data is essential for the development of a machine learning model in a subsequent phase of the project.

Personal data is stored and analyzed in pseudonymized form. Access to the pseudonymous key, which is stored on password-protected hardware, is limited to authorized researchers. The key is only used for decryption in the event of a participant's request for data deletion upon withdrawal.

3.3.19. Ethics

The ethical soundness of this study is substantiated by the ethics committee of the LBI DHP. The committee's approval to ethical standards and guidelines, is documented in Appendix C for reference and transparency. This underscores the study's commitment to upholding ethical principles and ensures the protection of the rights and well-being of the participants involved.

3.4. Audio and Pulse Synchronization

The pulse synchronization with the audio files has to be done manually for each participant. Figure 3.5 below describes the synchronization process flow.

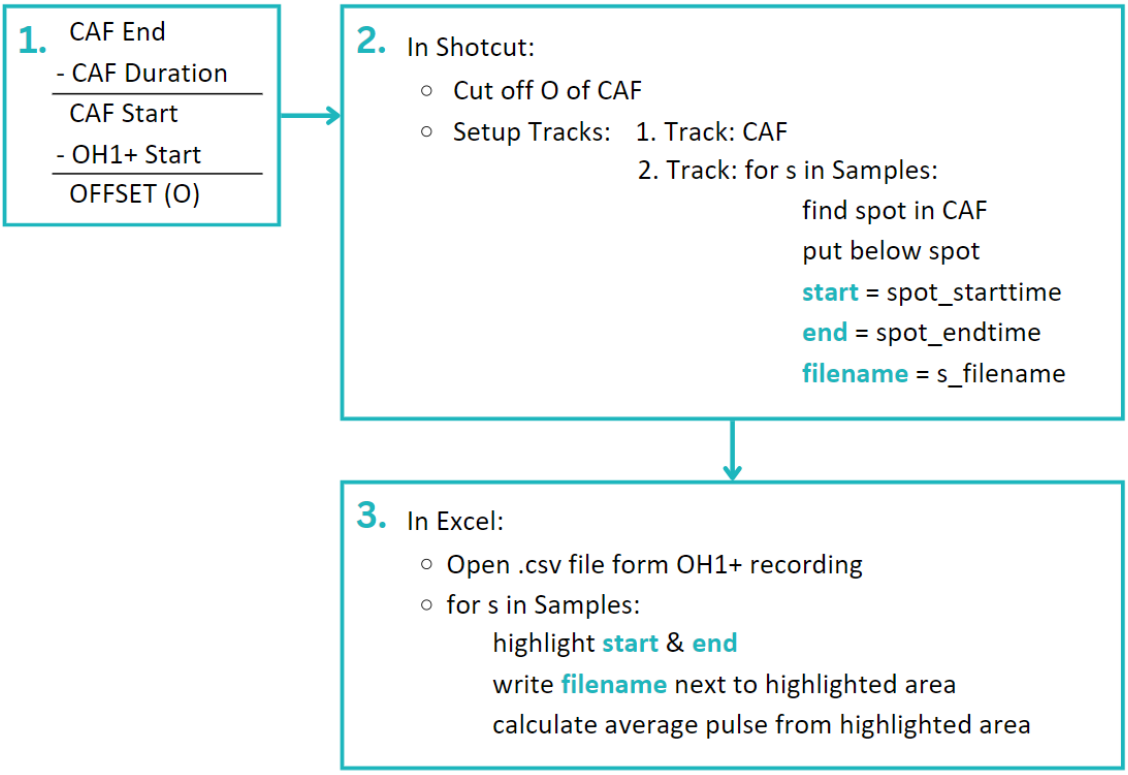


Figure 3.5.: Pulse and audio sample synchronization

Step 1: Since two smartphones are used to record the voice samples including one which records continuously (Continuous Audio File (CAF)), the synchronization of the pulse with the single voice samples can be done by making use of the UTC+1 time of the devices. The first step here is to examine the continuous voice recordings and their recording timestamps. The date the recording was saved and the duration are derived from the CAF. Furthermore, the pulse data is collected using the Polar Beat app, which provides the option to export a .csv file containing time-stamped pulse measurements at one-second intervals during the workout. The CAF ending time (safe time) and duration are then used to calculate the appropriate starting time of the voice recording. The offset of the audio file and the polar pulse recording are computed by subtracting the polar pulse starting time from the voice recording starting time, since the voice recording is started prior to the activation of the pulse recording.

Step 2: In the next step the voice recording is imported into Shotcut [21] which offers a range of video and audio editing features. The calculated offset is removed to achieve synchronization with the polar pulse recording and each individual cut audio sample is synchronized in a separate audio track with the CAF. This is done manually for each audio sample by having a look at the creation times.

Step 3: In the last step the starting and ending time of each audio sample are documented within the polar pulse recording, along with the corresponding file name, to provide crucial information about the specific timing and associated audio file. The average pulse of each audio file is calculated by taking the average of the pulse values between the starting and ending time.

4 out of 20 CAFs were recorded with an iOS device instead of using an Android device. Apparently, iOS drops the timestamps of voice recordings after 24 hours. If the files are not directly transferred to another device within this time frame, there is no way to synchronize the pulse with the single samples. This is a known issue of voice recordings among the Apple community, however, the conductor was initially unaware of this behavior. For further research, studies and data collection, it is important to keep this in mind.

3.5. Machine Learning-Based Intensity Classification

3.5.1. Manual Pre-Processing

The audio files used for machine learning are either recorded by the ***aktivtalk*** application (Method A and B) or recorded by a voice recording application that is installed on Smartphone 1 (Method C). These recordings are systematically named after the participant identifier (Txx), the method identifier y (yes, not sure, no), b

(BORG) or c (conductor) including an incremental number and the current intensity stage selected by the user. An example would be "T16c1-stageYes". Method C results in a single, comprehensive audio file that continuously records during this method, encompassing all data readings and self-assessed intensity zones. For this reason, the files from Method C are segmented into smaller audio files and accordingly relabeled.

It is observed that the files from Method A and B do not have the exact same length in seconds although the recording length is set to 15 seconds in the ***aktivtalk*** application. However, the length only ranges from about 14:00 seconds to 14:20 seconds and most files have a length of 14:14 seconds. Therefore, the files are all cut to 14:14 seconds.

Furthermore, a .csv file with all cut files and labels has been created. Initially, this .csv file features a classification scheme consisting of three discrete labels: 2 (light intensity), 1 (medium intensity) and 0 (high intensity). However, due to the challenges associated with predicting these three labels because of their similarities in the mel-spectograms, another labeling structure is added. Consequently, files characterized by light and medium intensity have been uniformly assigned the label "1", while files associated with high intensity have been uniformly designated with the label "0". For patients with cardiovascular diseases, it is crucial to train in a light to medium intensity area, and therefore it is decided to take a binary classification system into account as well.

3.5.2. Automated Pre-Processing

In order to represent the audio files in a more descriptive manner, mel-spectograms have been employed. These mel-spectograms serve as a visualization tool for mapping the relationship between frequency and time. The frequency is converted to the mel-scale which is declared as "a perceptual scale of pitches judged by listeners to be equal in distance from one another" by the University of California [26]. Using the mel-spectogram instead of a spectogram is a way of giving models sound data that is comparable to what a human would hear [9].

Down below the mel-spectograms for the three different intensity zones can be seen.

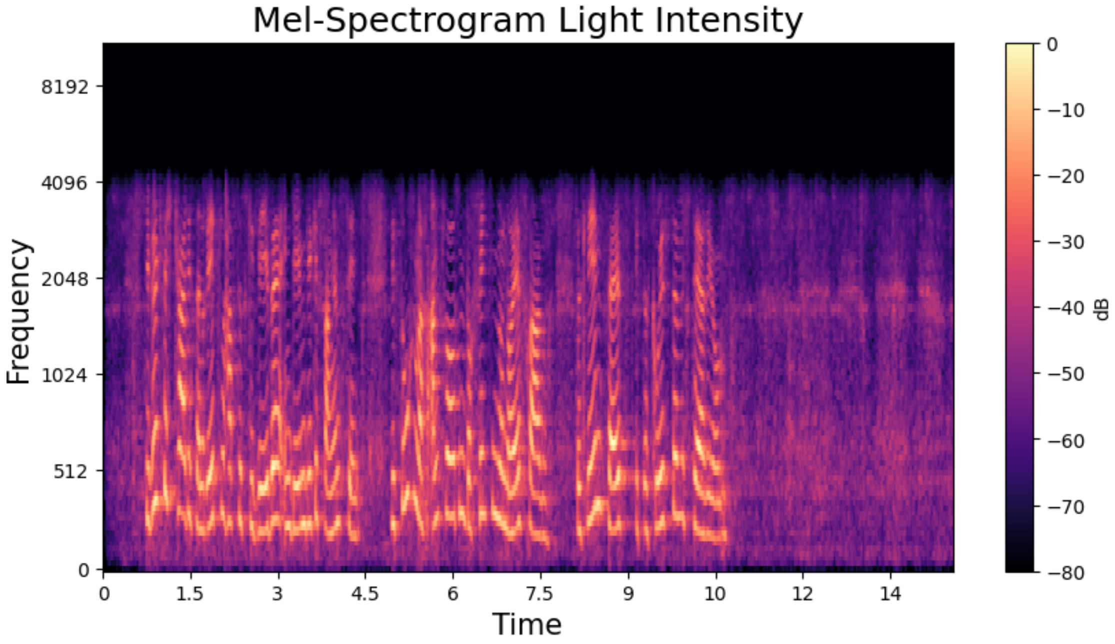


Figure 3.6.: Example of mel-spectogram of light intensity

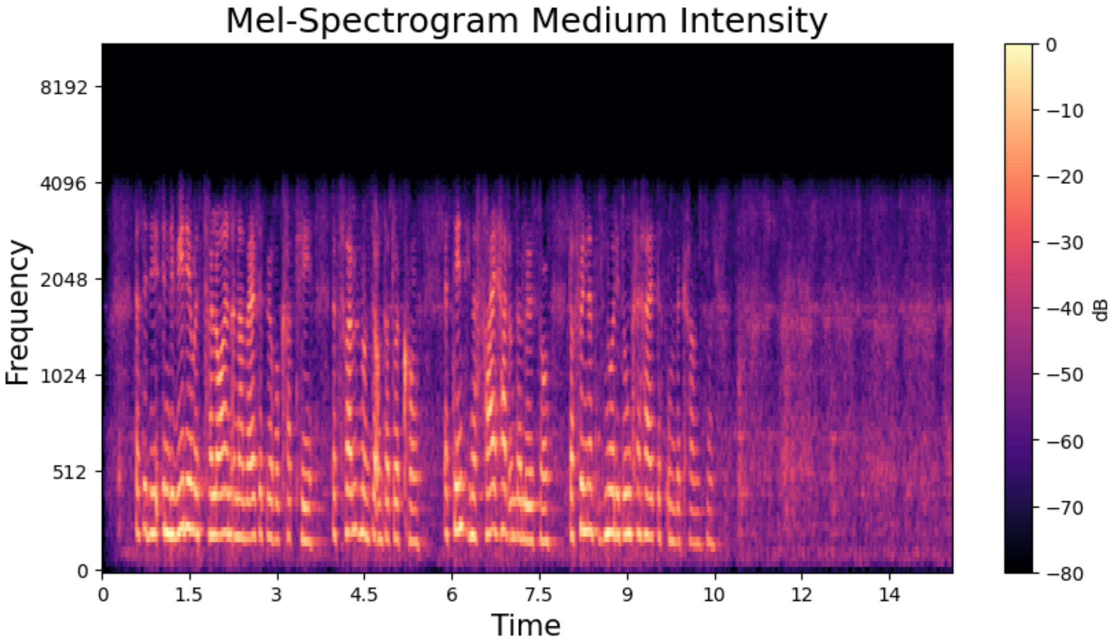


Figure 3.7.: Example of mel-spectogram of medium intensity

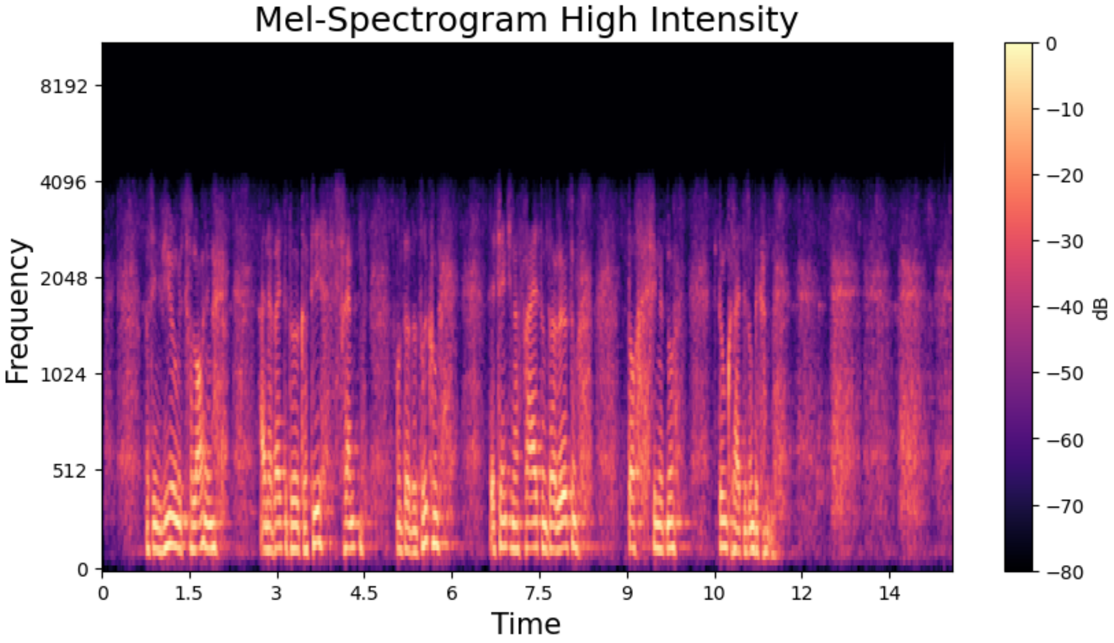


Figure 3.8.: Example of mel-spectogram of high intensity

The differences in the mel-spectograms in Figures 3.6, 3.7 and 3.8 are not obvious, however, it can be observed that in Figure 3.6 at around 4.5 seconds there is a small gap. In the mel-spectogram this could either be a breath or a speaking pause between sentences or commas. In Figure 3.7 and 3.8 these gaps occur more often, which could be an indicator for increasing intensity. Moreover, the yellow waves in Figure 3.7 appear more closer than in Figure 3.6. This could also be another indicator for increasing stress in speech. When looking at Figure 3.6 and 3.7 with light to medium intensity, they seem to be more similar than to Figure 3.8 with high intensity which makes it harder to distinguish from files with light to medium intensity. Nonetheless, it is crucial to note that these visual patterns may not uniformly apply to all individual files, as there may exist outliers or borderline cases warranting distinct analytical consideration.

Furthermore, the arithmetic mean of the transposed mel-spectrogram is computed. The transpose of the matrix is used in order to swap the time and frequency axes. This is helpful for treating the time frames as individual observations and the frequency components as features. The mean of each frequency bin is computed in order to reduce the dimension which helps to reduce the impact of short-term variations and noise in the audio signal. As a result, the feature representation may be more resistant to changes in the input signal. The relevant code implementation is provided in the subsequent listing 3.1:

```python
def get_features(df_in):
    extracted_features_self=[]
    extracted_features_pulse=[]
    for index in range(0,len(df_in)):
      filename = df_in.iloc[index]['filename']
      label_self = df_in.iloc[index]['label_3c']
      label_pulse = df_in.iloc[index]['label_pulse_3c']
      y, sr = librosa.load('/Study/'+filename,sr=28000)

      mfccs_features = librosa.feature.mfcc(y=y, sr=sr, n_mfcc=100)
      mfccs_scaled_features = np.mean(mfccs_features.T,axis=0)

      extracted_features_self.append([mfccs_scaled_features,label_self])
      extracted_features_pulse.append([mfccs_scaled_features,label_pulse])
    return(extracted_features_self,extracted_features_pulse)

X_self, X_pulse=get_features(metadata)
```

```
extracted_features_self=pd.DataFrame(X_self,columns=['feature','class'])
extracted_features_pulse=pd.DataFrame(X_pulse,columns=['feature','class'])
```

Listing 3.1: Feature extraction

Furthermore, for computing the features of the single audio samples, an alternative *get_features* function is proposed in order to generate a higher number of samples. The idea is to "supersample" the single audio files by segmenting them into smaller conjunct pieces. Instead of having one large audio file of 14 seconds, 4 audio files of 5 seconds are generated as follows in listing 3.2:

```
def get_features(df_in):
    extracted_features_self=[]
    extracted_features_pulse=[]
    for index in range(0,len(df_in)):
      filename = df_in.iloc[index]['filename']
      label_self = df_in.iloc[index]['label_3c']
      label_pulse = df_in.iloc[index]['label_pulse_3c']
      y, sr = librosa.load('/Study/'+filename,sr=28000)

      duration = 5.0
      offset = 0.0
      for i in range(0,11,3):
        y, sr = librosa.load(path_to_file+filename,sr=28000, offset=offset+i,
    duration=duration)
        mfccs_features = librosa.feature.mfcc(y=y, sr=sr, n_mfcc=100)
        mfccs_scaled_features = np.mean(mfccs_features.T,axis=0)

        extracted_features_self.append([mfccs_scaled_features,label_self])
        extracted_features_pulse.append([mfccs_scaled_features,label_pulse])
    return(extracted_features)

X=get_features(metadata)
extracted_features_df=pd.DataFrame(X,columns=['feature','class'])
```

Listing 3.2: Feature extraction with supersampling

The supersampling idea can be described with the following Figure 3.9:

As mentioned in Subsection 3.5.1 the dataset was relabeled in order to only have samples with light to medium intensity (1) or high intensity (0). This results in an imbalance of the number of samples in the target classes. For this reason an upsam-

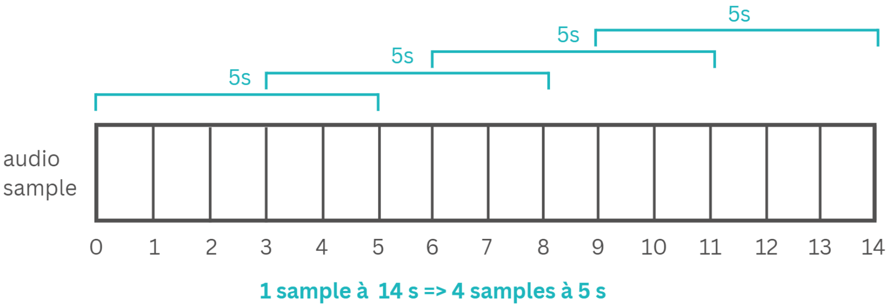


Figure 3.9.: Supersampling

pling technique called SMOTEENN (Synthetic Minority Over-sampling Technique combined with Edited Nearest Neighbors) [22] is used in order to generate more samples in the 0 class since it is underrepresented. This helps to create a balanced dataset with an equal amount of samples in both target classes.

3.5.3. Model

The model is defined as a Sequential Neural Network, which means that layers are added one after the other in a linear way (see listing 3.3). The model starts with an input layer with 559 units. The input shape is set to (100,), which means the input data is expected to have 100 features. After each dense layer, a Rectified Linear Unit (ReLU) activation function is applied. ReLU is a common choice for activation functions in hidden layers, as it introduces non-linearity into the model. Furthermore, the model consists of multiple hidden layes with a decreasing number of units to reduce the dimension of the data gradually. The last dense layer has 2 units and a sigmoid activation function which is commonly used in binary classification problems. When predicting 3 labels a softmax activation function of the last layer and a categorical_crossentropy loss function is used instead.

```
kfold = KFold(n_splits=5, shuffle=True)

fold_no = 1
for train, test in kfold.split(inputs, targets):
  num_labels=2
```

```
model=Sequential()
model.add(Dense(559,input_shape=(100,)))
model.add(Activation('relu'))
model.add(Dropout(0.5))
model.add(Dense(280,input_shape=(100,)))
model.add(Activation('relu'))
model.add(Dropout(0.3))
model.add(Dense(140,input_shape=(100,)))
model.add(Activation('relu'))
model.add(Dropout(0.3))
model.add(Dense(120,input_shape=(100,)))
model.add(Activation('relu'))
model.add(Dense(60,input_shape=(100,)))
model.add(Activation('relu'))
model.add(Dense(30,input_shape=(100,)))
model.add(Activation('relu'))
model.add(Dense(15,input_shape=(100,)))
model.add(Activation('relu'))

model.add(Dense(num_labels))
model.add(Activation('sigmoid'))

model.compile(loss='binary_crossentropy',metrics=['accuracy'],
  optimizer='adam')

num_epochs = 200
num_batch_size = 64
history = model.fit(inputs[train], targets[train], batch_size=
  num_batch_size, epochs=num_epochs, validation_data=(inputs[test],
  targets[test]), verbose=1)
```

Listing 3.3: Sequential neural network with 5-fold validation

Cross-validation is used in order to assess the performance and generalization capabilities of the machine learning model. It helps to estimate how well the model is likely to perform on unseen data. The dataset is divided into 5 folds with each one training and testing on different subsets. The risk of overfitting can be reduced by training and testing on multiple folds. Moreover, cross-validation makes a more efficient use of the data, since it allows generating multiple performance estimates. These estimates are averaged in this specific case in order to obtain a more reliable estimate of the model's performance that is less sensitive to the particular data partitioning.

4. Results

4.1. Systematic Literature Review

4.1.1. Selected Studies

On Google Scholar and Springer there is only a full-text search available and therefore 1260 papers were found on Google Scholar and 51 on Springer. Because of the exclusion criteria, 9 starting, 12 forward and 11 backward paper were found in Google Scholar. In Springer 3 starting, 3 forward and 6 backward paper were identified. Duplicates were removed and three more sources turned out to be unsatisfactory because two were not freely accessible and another two were about the healthy population. After that, 7 paper were included, of which 4 were studies. 2 more studies were added after a manual search at the end. The studies were no older than 2014.

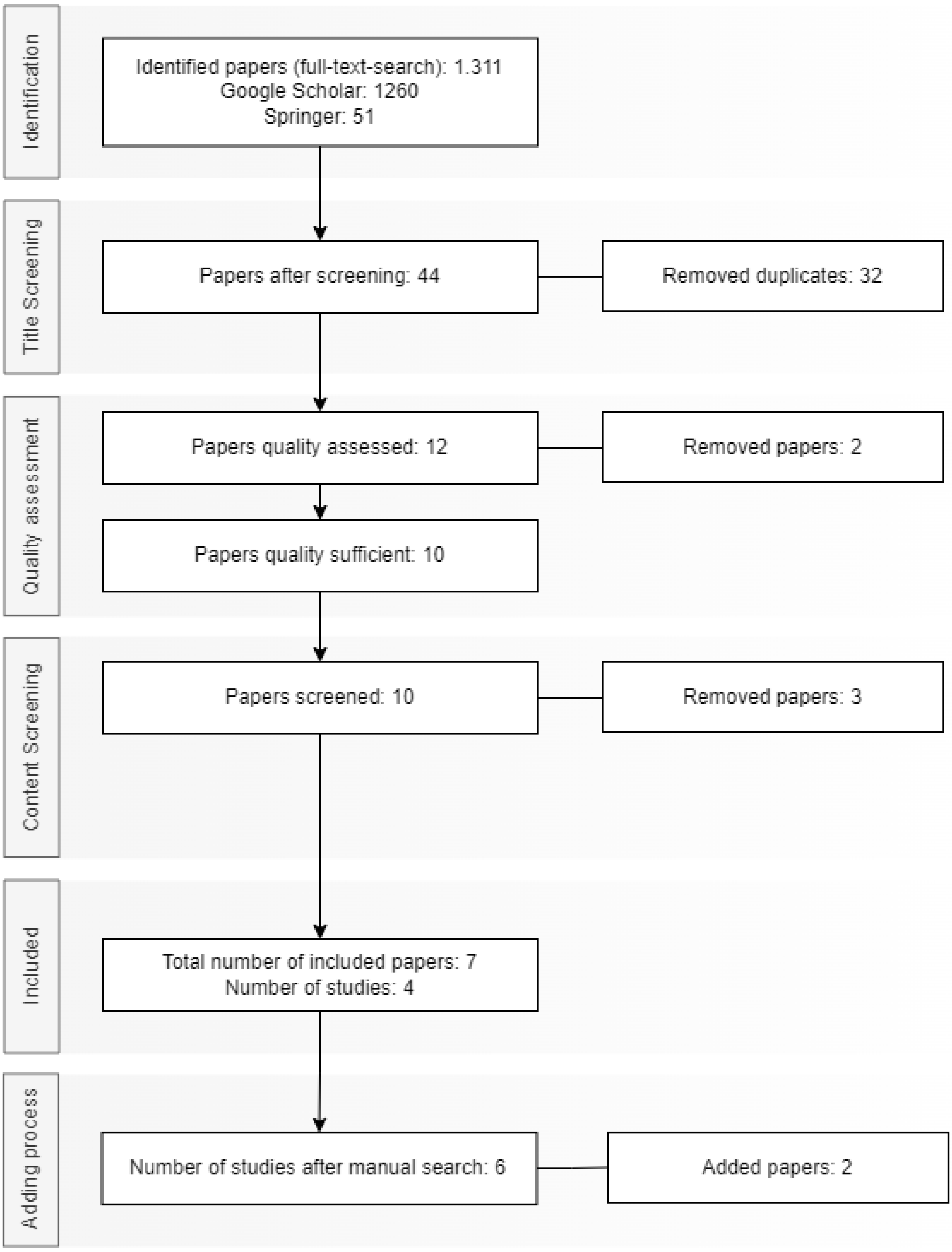


Figure 4.1.: Flowchart of the Systematic Literature Review

Paper	Title	Year	Author	Mode of activity	Protocol	Quality
1	Modified Talk Test: a Randomized Crossover Trial Investigating the Comparative Utility of Two "Talk Tests" for Determining Aerobic Training Zones in Overweight and Obese Patients	2021	Orizola-Cáceres et al.	Incremental cycloergometric tests	10-min warm-up increments every 3 min last 30s of each stage 40 words + TT	Q1/Q1
2	Reliability of graded cycling test with talk test and 30-s chair-stand test in men with prostate cancer on androgen deprivation therapy	2020	Aabo et al.	Graded cycling test	2-min warm-up (15 W, 60 rpm) increments every min (15 W) last 10s of each stage 30 words + TT	Q2
3	Validity of the Talk Test as a Method to Estimate Ventilatory Threshold and Guide Exercise Intensity in Cardiac Patients	2020	Sørensen et al.	Submaximal test on cycle ergometer	2-min warm-up (0 W, 60 rpm) increments every min (15 W) last 10s of each stage 30 words + TT	Q2/Q2/Q2
4	"Graded Cycling Test with Talk Test" Is a Reliable Test to Monitor Cardiovascular Fitness in Patients with Minor Stroke	2017	Krawcyk et al.	Graded cycling test	2-min warm-up (15 W, 60 rpm) increments every min (15 W) end of each stage 30 words + TT	Q2/Q2/Q2/Q2
5	Graded Cycling Test Combined With the Talk Test Is Responsive in Cardiac Rehabilitation	2016	Nielsen and Vinther	Graded cycling test	2-min warm-up (15 W, 60 rpm) increments every min (15 W) end of each stage 30 words + TT	Q1/Q2/Q2
6	The Graded Cycling Test Combined With the Talk Test Is Reliable for Patients With Ischemic Heart Disease	2014	Nielsen et al.	Graded cycling test	2-min warm-up (15 W, 60 rpm) increments every min (15 W) end of each stage 30 words + TT	Q1/Q2/Q2

Table 4.1.: Characteristics of the included studies

4.1.2. Comparison of Studies

Five of the six papers followed the same testing procedure (see Table 4.1). The TT was done on a cycle ergometer and there was a 2-min warm-up with 15 W and 60 rpm in three cases. Sørensen et al. [23] did a warm-up with 0 W. After the warm-up the load was increased by 15 W every minute. At the end of each minute, the TT was done by reciting 30 words. The paper by Orizola-Cáceres et al. [16] included a 10-minute warm-up and 3-minute increments. A 40-word paragraph was recited in the last 30s of each stage followed by the TT. The workload was not described in this paper.

The mode of activity is described with different terms in the papers. However, it can be assumed that they have a strong correlation because the protocols are very similar or equal.

Table 4.2 shows the diagnoses of the participants. Paper 1 [16] only includes patients who are diagnosed with overweight/obesity. Papers 3, 5 and 6 [23, 15, 14] focus on cardiovascular diseases, whereas almost all patients in paper 4 [11] have lacunar strokes. Paper 2 [1] includes patients with various diagnoses. Even though 3 papers include mostly patients with cardiovascular diseases, only paper 6 [14] contains a representative amount of patients treated with beta-blockers (72%). In paper 5 [15] only one patient takes beta-blockers, whereas in paper 3 [23] the amount of patients treated with beta-blockers is not mentioned.

Furthermore, the duration of the corresponding studies can be seen in Table 4.2. Most studies take place on one day, even if two tests are performed. The duration of paper 5 [15] is 8 weeks since the pretest is done before the rehabilitation and the posttest takes place after the 8-week rehabilitation. Paper 5 [15] is the only study about whether the TT is able to detect a change in the physical fitness of the patients after a certain period.

In five out of the six studies examined, an imbalance in gender distribution was observed. This disparity in participant gender may introduce a potential bias, affecting the generalizability of the findings. For this reason the study on the ***aktivtalk*** is done by maintaining a balanced gender distribution among participants, contributing to a more representative and inclusive research sample.

Paper	Participants diagnosis	Age in yrs	Gender ratio	Metric Properties	Duration	Reliability
1	Overweight/obese	34.9 +- 6.7	6f/13m	TTT power output agreements: "first no" and VT2 "last yes" and VAS 4-5 of ATT ATT power output agreements: VAS 2-3 and VT1 VAS 6-7 and VT2	1 day	-
2	16 cardiovascular diseases, 5 diabetes, 29 hypertension, 27 dyslipidemia, 3 osteoporosis, 10 castration-resistant prostate cancer	70.8 +- 6.6	60m	Test 1: 122.8 ± 33.1 W Test 2: 120.3 ± 32.6 W Test 1 and 2 $\text{ICC}_{2.1}$ 0.90 (0.84-0.94) for GCT-TT (W)	1 day	Excellent
3	8 coronary artery bypass graft, 6 percutaneous coronary intervention, 4 heart valve surgery, 2 heart failure	65 ± 8.5	2f/18m	TT_{pos}: 77 ± 24 W TT_{eq}: 101 ± 27 W TT_{neg}: 116 ± 27 W VT: 107 ± 39 W	1 day	-
4	60 lacunar stroke, 45 hypertension, 7 diabetes, 10 arthritis	67 (44-85)	19f/41m	Test 1 Test 2 TT+ 88.5 ± 37.2 W 88.8 ± 36.4 W TT- 114.8 ± 37.0 W 114 ± 35.6 W TT+ $\text{ICC}_{2.1}$ 0.92 (0.87-0.95) TT- $\text{ICC}_{2.1}$ 0.97 (0.95-0.98)	1 day	Excellent
5	81 ischemic heart disease, 16 coronary artery bypass graft, 29 percutaneous coronary intervention, 13 stable angina, 8 heart valva surgery, 2 chronic obstructive pulmonary disease, 4 other	63.3 +- 9.7	25f/68m	Pretest: 104.3 ± 30.4 W Posttest: 122.3 ± 37.0 W	8 weeks	-
6	ischemic heart diseases: postoperative coronary artery bypass graft, percutaneous intervention, stable angina pectoris, heart failure	36 - 82	30f/34m	Test 1 Test 2 TT+ 61.0 ± 27.6 64.2 ± 28.5 TT- 94.4 ± 30.1 96.2 ± 28.4 TT± 82.9 ± 29.1 84.1 ± 28.4 Rater 1 86.1 ± 24.0 88.5 ± 26.1 Rater 2 89.3 ± 23.6 87.8 ± 23.5 TT+ ICC: 0.90 (0.84-0.94) TT- ICC: 0.90 (0.83-0.94) TT± ICC: 0.91 (0.84-0.95) Rater 1 ICC: 0.81 (0.70-0.88) Rater 2 ICC: 0.88 (0.81-0.93)	1 day	Good/ Excellent

Table 4.2.: Detailed characteristics of the included studies

Not all studies use the same measurement metrics (see Table 4.2). Papers 3, 4 and 6 [23, 11, 14] measure the power output of the watts regarding the positive and negative stages of the TT. The mean of the watts of the positive stage (test 1) in paper 4 [11] is more than 10 watts higher than for paper 3 [23]. The study duration, the protocol and the age of the participant are around the same for both studies. However, study 4 [11] includes more participants and almost all are diagnosed with lacunar strokes. Study 3 [23] includes patients with cardiovascular diseases and almost all participants are men. The difference in the diagnosis, gender and the number of participants could be the reasons for the difference in the mean. In study 6 [14] the mean of the watts of all stages is significantly lower when compared to studies 3 and 4 [23, 11]. In Table 4.2 it can be seen that the gender distribution is almost balanced and the mean of the age was not specified. It can also be observed that the raters measured the workload at which the patients were no more able to speak comfortably a bit earlier. There was a difference when measuring this point in time for the two raters in test 1. However, in test 2 the observations of the two raters were more similar [14].

In studies 1 and 3 [16, 23] the VT was measured against the stages of the TT. Agreements were found for the ATT and TTT between the stages of the tests and the VT1/VT2 in paper 1 [16]. However, in study 3 [23] the VT lies between the $\mathrm{TT}_{\mathrm{eq}}$ and $\mathrm{TT}_{\mathrm{neg}}$ and is not significantly different from them. However, there is no clear agreement. Study 3 [23] concludes that the TT can not be used as a surrogate for the VT. Previous studies have also shown inconsistencies in the relationship between the TT and the VT.

4.2. Digital Talk Test

In documenting the ***aktivtalk*** app within the arc42 framework, it is important to note that the prototype's documentation is not entirely complete. Prototypes, including the ***aktivtalk*** app, are designed to demonstrate and validate concepts and functionalities in a fast way by prioritizing efficiency over exhaustive documentation. The goal of the prototype is to quickly prove the concept. As the ***aktivtalk*** app improves, documentation can be expanded in line with arc42 principles, allowing for a more comprehensive and refined architectural documentation.

RQ2 is directly addressed by using the Talk Test procedure and the prototype explained with the arc42 framework in the sections below.

4.2.1. Introduction

The ***aktivtalk*** application is a prototype of the digital version of the Talk Test. The prototype is developed as a mobile application which helps users to self-assess their current exercise intensity level during workouts. The assessment type, repetition and testing interval can be chosen by the user. After starting the workout, the user has the option to do the Talk Test by only using speech. After reading out loud the user can either assess their exercise intensity zone through Speech Recognition or haptically giving the responsive answer.

Stakeholder

- LBI DHP: The interest of the institute lies in supporting cardiovascular diseases patients during rehabilitation.
- CVD patients: Patients with cardiovascular diseases require methods other than relying on the pulse to assess their exercise intensity, especially due to the use of specific treatment medications like beta blockers, which lower the pulse.

4.2.2. Quality Goals

- High usability and user experience: The primary objective of this goal is to ensure that the ***aktivtalk*** application offers an exceptional user experience by being easy-to-navigate and intuitive. The application feature an intuitive interface with consistent design elements, minimizing the learning curve for users.

- Reliable recording of voice samples: For the prototype it is essential that the voice recordings are being collected and stored in an appropriate manner.
- Reliable Talk Test procedure: During the workout it is important that the procedure is reliable, since it would interfere with the study.

4.2.3. Constraints

- Prototype development: As of its current state, the ***aktivtalk*** mobile application exists as a prototype, and it has not been deployed to end-users. This constraint implies that the application is not yet available for widespread use, limiting its accessibility and availability to the broader user base.
- Execution limited to source code: To interact with the app, users are required to execute and test the application by running its source code in an emulator or on a physical device.
- Platform specificity - Nokia 7.2 and Android 14.0: This application has undergone primary development and testing specifically on Nokia 7.2 devices utilizing the Android operating system.
- Schedule and time constraints: Due to the schedule of the master thesis the prototype was developed with time constraints, since the conduction of the user study relied on the availability of the ***aktivtalk*** application.

4.2.4. Context and Scope

The ***aktivtalk*** application is developed as an independent application. However, the future goal is to integrate it into the LBI DHP's ***aktivplan*** application which is specifically designed to support patients with cardiovascular diseases during rehabilitation. Since patients are often required to engage in home-based exercise without any form of external monitoring, there is a need for a user-friendly and cost-efficient tool to address the evaluation of exercise intensity.

4.2.5. Solution Strategy

GUI Prototype

The development will prioritize an intuitive and easy-to-navigate interface in order to achieve a high usability and user experience. The initial iteration of the prototype was completed before the first in-person meeting at LBI DHP in Salzburg. During the meeting improvements and ideas about the prototype were discussed. The final prototype can be viewed in Figure 4.2 and Figure 4.3, or by visiting the following link:`https://www.figma.com/file/dww2PGcSY4VT2l8kbibpzP/aktivtalkprototypegeiger?type=designnoid=0%3A1t=JHn8ZWAiLdPRunIq-1`

Reliable behavior

For the reliable recording of voice samples, the prototype will implement a robust data storage mechanism. Each voice sample will be continuously recorded, labeled with the participants identification number, id, method and current exercise intensity stage. Figma will again be utilized to create interactive prototypes, allowing stakeholders to simulate the Talk Test procedure and provide valuable insights before the actual implementation.

Furthermore, the reliable recording of the voice samples and the seamless procedure is guaranteed by extensive manual testing before conducting the study. Furthermore, a study test trial ensures that the ***aktivtalk*** prototype is working correctly.

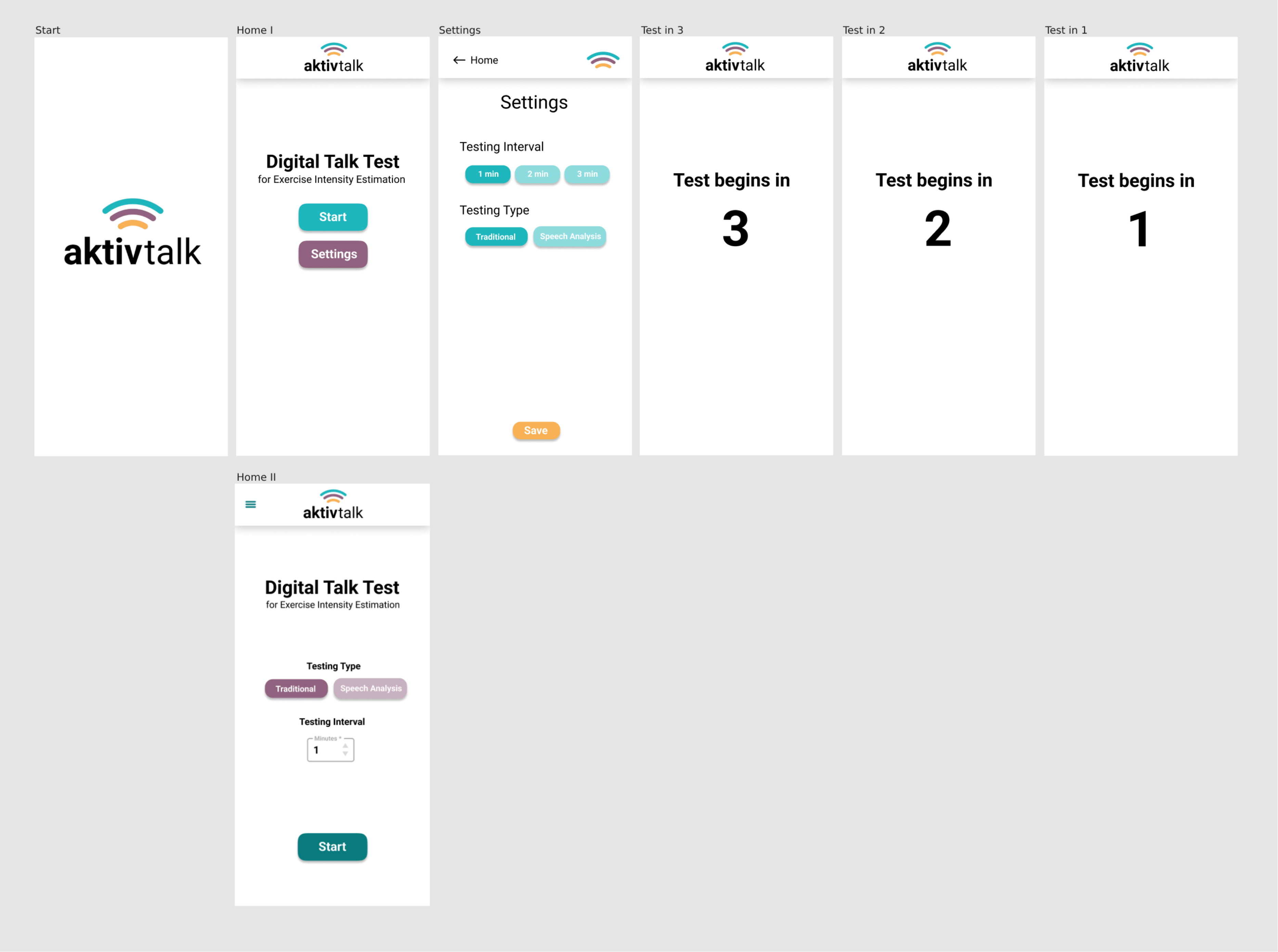
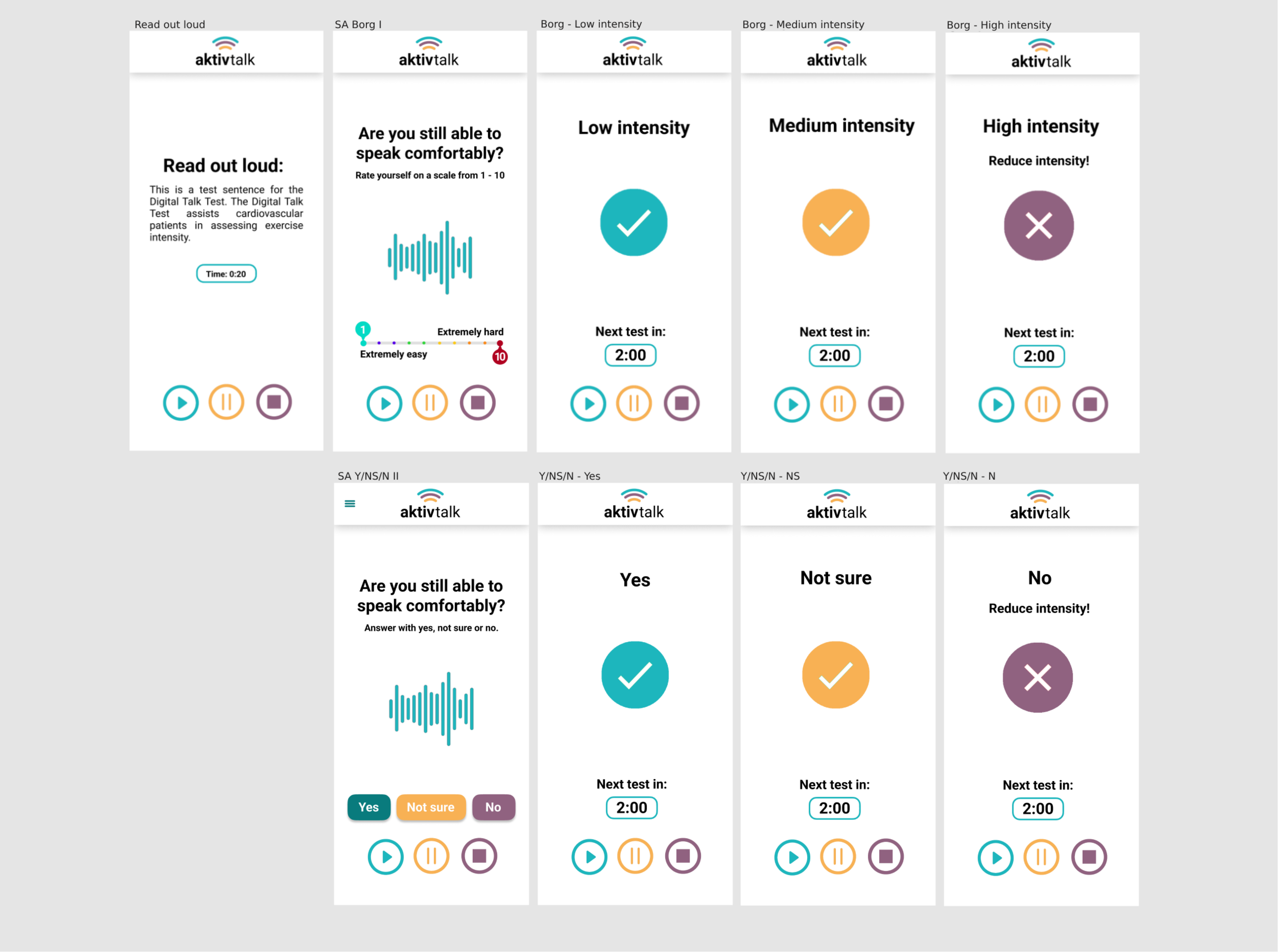


Figure 4.2.: Prototype part I



Read out loud

Read out loud:

This is a test sentence for the Digital Talk Test. The Digital Talk Test assists cardiovascular patients in assessing exercise intensity.

Time: 0:20

SA Borg I

Are you still able to speak comfortably?

Rate yourself on a scale from 1 - 10

Borg - Low intensity

Low intensity

Next test in: 2:00

Borg - Medium intensity

Medium intensity

Next test in: 2:00

Borg - High intensity

High intensity

Reduce intensity!

Next test in: 2:00

SA Y/NS/N II

Are you still able to speak comfortably?

Answer with yes, not sure or no.

Y/NS/N - Yes

Yes

Next test in: 2:00

Y/NS/N - NS

Not sure

Next test in: 2:00

Y/NS/N - N

No

Reduce intensity!

Next test in: 2:00

Figure 4.3.: Prototype part II

4.2.6. Use Cases

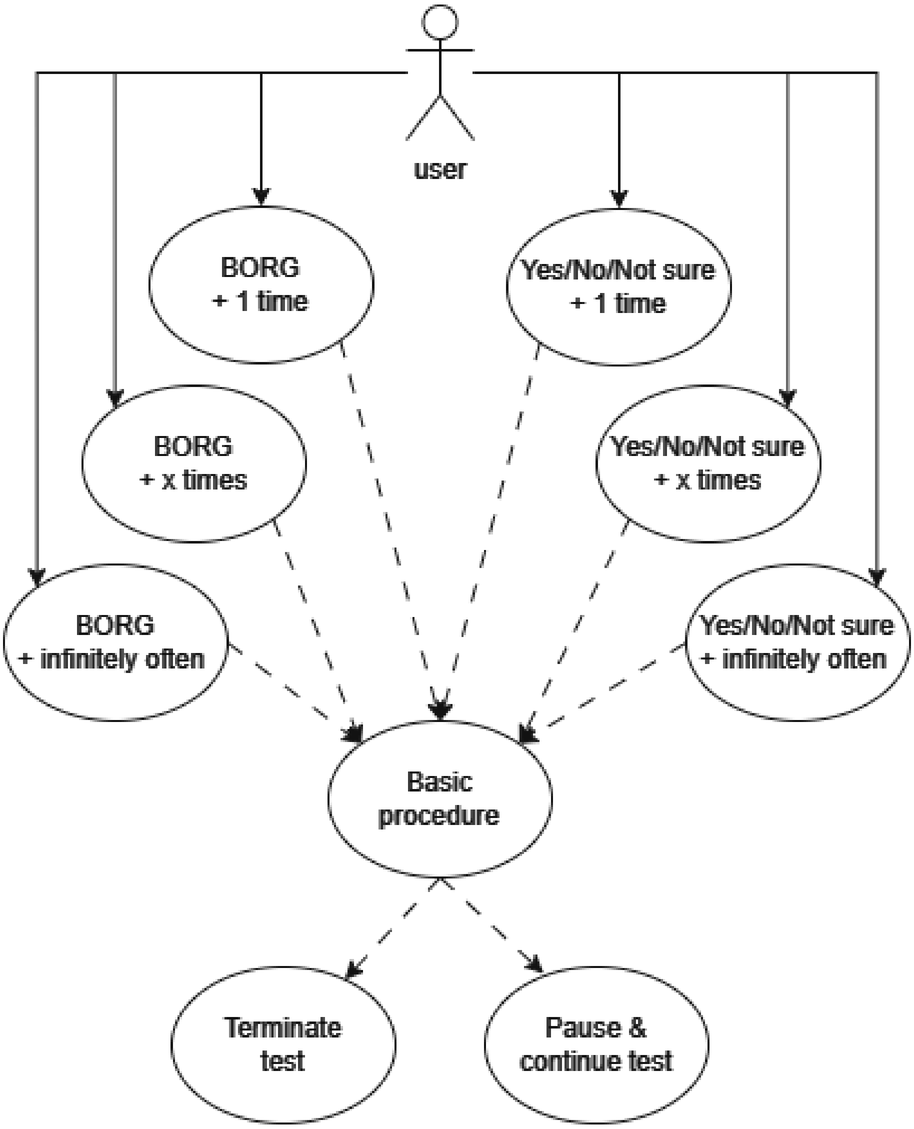


Figure 4.4.: Use Case diagram

UC1.1: BORG + 1 time

Name	BORG + 1 time
ID	UC1.1
Description	The user chooses the Borg scale assessment and a single test repetition interval.
Precondition	• The user is in the home view.
Procedure	• The user presses the "Borg scale" button in order to select the assessment type. • The user chooses "Once" in the repetition interval dropdown menu.
Postcondition	• The "Borg scale" button is active (dark purple background) and the "Yes/Not sure/No" button is inactive (light purple background). • The repetition interval's dropdown menu displays "Once".
Alternative	-

UC1.2: BORG + x times

Name	BORG + x times
ID	UC1.2
Description	The user chooses the Borg scale assessment and a testing interval of x.
Precondition	• The user is in the home view.
Procedure	• The user presses the "Borg scale" button in order to select the assessment type. • The user chooses "x times" in the repetition interval dropdown menu.
Postcondition	• The "Borg scale" button is active (dark purple background) and the "Yes/Not sure/No" button is inactive (light purple background). • The repetition interval's dropdown menu displays "x times".
Alternative	-

UC1.3: BORG + infinitely often

Name	BORG + infinitely often
ID	UC1.3
Description	The user chooses the Borg scale assessment and an infinite testing interval.
Precondition	• The user is in the home view.
Procedure	• The user presses the "Borg scale" button in order to select the assessment type. • The user chooses "Infinitely" in the repetition interval dropdown menu.
Postcondition	• The "Borg scale" button is active (dark purple background) and the "Yes/Not sure/No" button is inactive (light purple background). • The repetition interval's dropdown menu displays "Infinitely".
Alternative	-

UC2.1: YNNS + 1 time

Name	YNNS + 1 time
ID	UC2.1
Description	The user chooses the Yes/Not sure/No assessment and a single test repetition interval.
Precondition	• The user is in the home view.
Procedure	• The user presses the ”Yes/Not sure/No” button in order to select the assessment type. • The user chooses ”Once” in the repetition interval dropdown menu.
Postcondition	• The ”Yes/Not sure/No” button is active (dark purple background) and the ”Borg Scale” button is inactive (light purple background). • The repetition interval's dropdown menu displays ”Once”.
Alternative	-

UC2.2: YNNS + x times

Name	YNNS + x times
ID	UC2.2
Description	The user chooses the Yes/Not sure/No assessment and a testing interval of x.
Precondition	• The user is in the home view.
Procedure	• The user presses the "Yes/Not sure/No" button in order to select the assessment type. • The user chooses "x times" in the repetition interval dropdown menu.
Postcondition	• The "Yes/Not sure/No" button is active (dark purple background) and the "Borg scale" button is inactive (light purple background). • The repetition interval's dropdown menu displays "x times".
Alternative	-

UC2.3: YNNS + infinitely often

Name	YNNS + infinitely often
ID	UC2.3
Description	The user chooses the Yes/Not sure/No assessment and an infinite testing interval.
Precondition	• The user is in the home view.
Procedure	• The user presses the "Yes/Not sure/No" button in order to select the assessment type. • The user chooses "Infinitely" in the repetition interval dropdown menu.
Postcondition	• The "Yes/Not sure/No" button is active (dark purple background) and the "Borg Scale" button is inactive (light purple background). • The repetition interval's dropdown menu displays "Infinitely".
Alternative	-

UC3: Basic Procedure

Name	Basic Procedure
ID	UC3
Description	The user is guided through the procedure of the Talk Test and assesses their exercise intensity.
Precondition	• The user is in the home view. • The user has successfully performed one of the following use cases: UC1.1., UC1.2, UC1.3 UC2.1, UC2.2., UC2.3
Procedure	• The user presses the ”Start” button in order to start the workout. • The countdown appears in visual and audible form. • The user is requested to read out loud a paragraph of approximately 30 words. • The user is requested to either assess themselves on a scale from 6 - 20 or answer with ”Yes”, ”No” or ”Not sure” to the question if talking was still comfortable either verbally or haptically.
Postcondition	• The user receives the feedback according to the previous self-asssessment. • The countdown appears and shows when the next test occurs.
User Alternatives	• The user chooses the wrong intensity by mistake. • The user can press on the ”Retake answer” button within 10 seconds in order to another assessment.
System Alternatives	• The app fails to receive the verbal self assessed intensity by the user. • The user can still submit their answer in a haptic form.

UC4: Terminate test

Name	Terminate test
ID	UC4
Description	The user terminates the Talk Test at some point during the workout.
Precondition	• The user has successfully performed UC3.
Procedure	• The user clicks on the purple terminate button in the bottom right corner. • The user confirms the dialogue with ”Yes” or ”Ja”.
Postcondition	• The user is in the home view again.
Alternative	-

UC5: Pause & continue test

Name	Pause & continue test
ID	UC5
Description	The user pauses and continues the Talk Test at some point during the workout.
Precondition	• The user has successfully performed UC3.
Procedure	• The user clicks on the orange pause button on the bottom left corner of the screen. • The pause button turns into a turquoise continue button. • The user clicks on the turquoise continue button.
Postcondition	• The user is in the same view as before pausing and continuing.
Alternative	-

UC6: Language Switch (German and English)

Name	Language Switch (German and English)
ID	UC6
Description	The user switches the language from german to english or vice versa.
Precondition	• The user is in the home view.
Procedure	• The user clicks on the austrian or british flag in the top right corner.
Postcondition	• The language changes from german to english or the other way around.
Alternative	-

4.2.7. Risks and Mitigation

Speech Recognition terminates after timeout

On Android, speech recognition terminates after the speaker pauses, due to a short timeout period. Depending on the device and Android OS version, the timeout seems to fluctuate. According to the author of the plugin, no pause exceeded more than five seconds.

speech_to_text [24], which offers speech recognition functionalities, begins immediately after the user is asked whether reading is still comfortable or not. If the user is not responding within the timeout, speech_to_text terminates. Therefore, a workaround, which involves invoking speech_to_text immediately after it terminates, is implemented. speech_to_text only stops re-invoking if the user's answer is given correctly.

Speech Recognition is not able to detect single numbers

The speech_to_text library sometimes detects the spoken numbers by the users as words and not as numbers. Therefore, for perceiving the user's response there should be no distinction made between numbers and words representing the answer. Furthermore, the speech_to_text plugin also encounters difficulty in recognizing a single word. Consequently, the user is encouraged to say "stage" before their respective answer.

Introducing haptic option to choose intensity

Since the Speech Recognition can cause problems, it is important to provide alternative options for communicating exercise intensity from the user's perspective. Therefore, haptic option like scales and buttons have been introduced in order to choose the exercise intensity when the Speech Recognition fails, is not available or simply not preferred.

4.2.8. High-Fidelity Prototype

The most crucial screens of the final Flutter [4] Prototype in Dart [6] are presented in Figure 4.5. This prototype includes a language switch feature for switching between German and English on the Home Screen. Users can also customize the warm-up duration and the number of repetitions for the Talk Test on this screen. Additionally, a countdown timer has been added before the Talk Test begins to enhance usability, allowing users time to prepare for the test. The selected exercise intensity is displayed along with the corresponding intensity zone. Users also have the option to retake their assessment within a 10-second window.

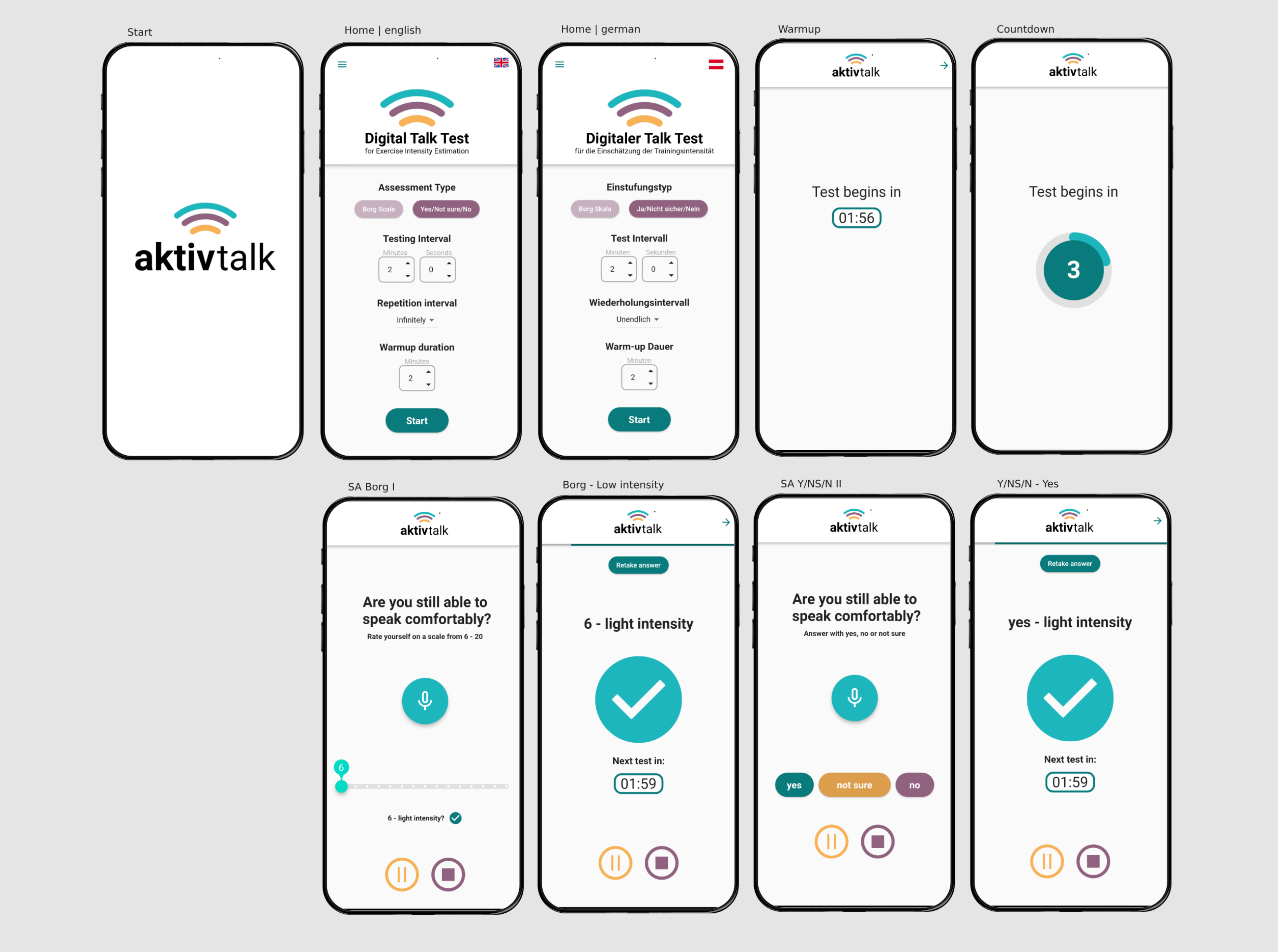


Figure 4.5.: Final flutter prototype

5. Evaluation

5.1. User Study Findings

5.1.1. Age and Gender

The 20 participants are distributed evenly across their self-identified gender. In total 10 females and 10 males took part in the study. Since the participants are recruited through the researcher's circle of acquaintances the age is not evenly distributed among various age groups. The mean age of the participants is 26 and the median is 24.5 (see Figure 5.1). The standard deviation has a value of 8.26 which indicates some degree of variability in the ages.

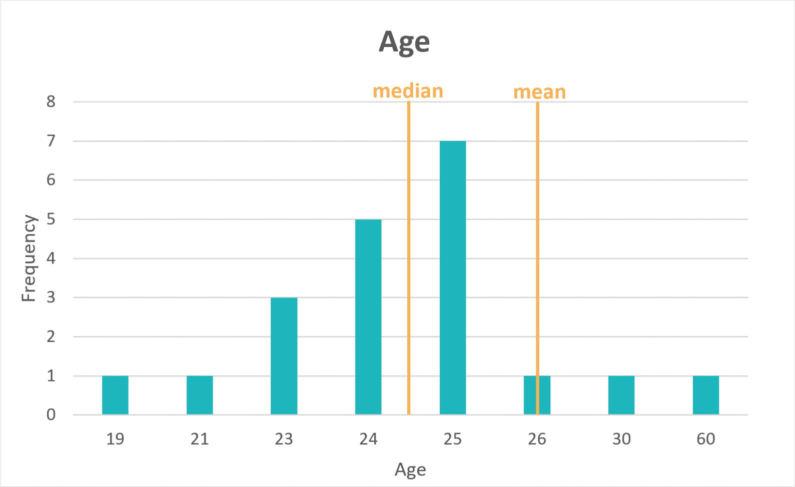


Figure 5.1.: Age

5.1.2. Physical Activity

The physical activity of participants is investigated by using a standardized questionnaire called Rapid Assessment of Physical Activity (RAPA). In total, 11 of the participants are active, 3 are under-active regular, 4 are under-active regular with light activities and 2 are under-active as shown in Figure 5.2.

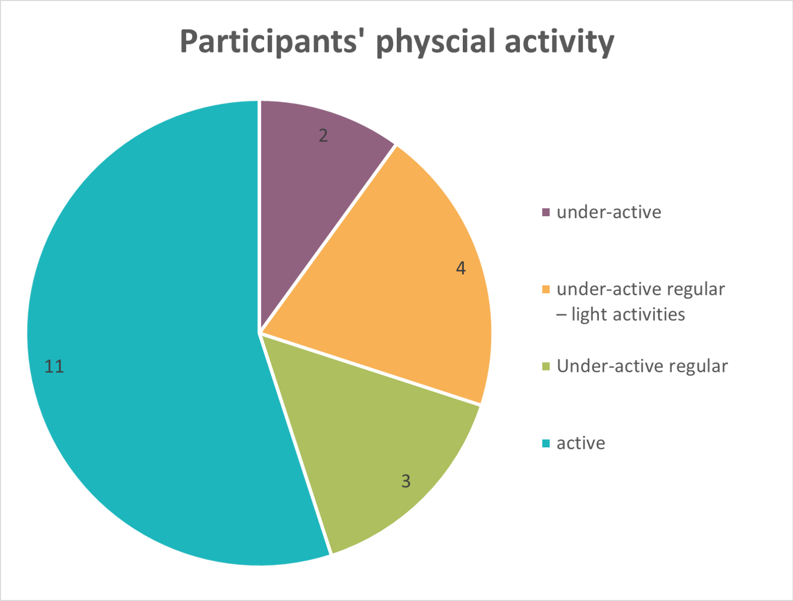


Figure 5.2.: Physical activity

The participants were also asked about the types of activities they regularly engage in. The most common activities were hiking, running, biking, tennis and weight lifting (see Figure 5.3).

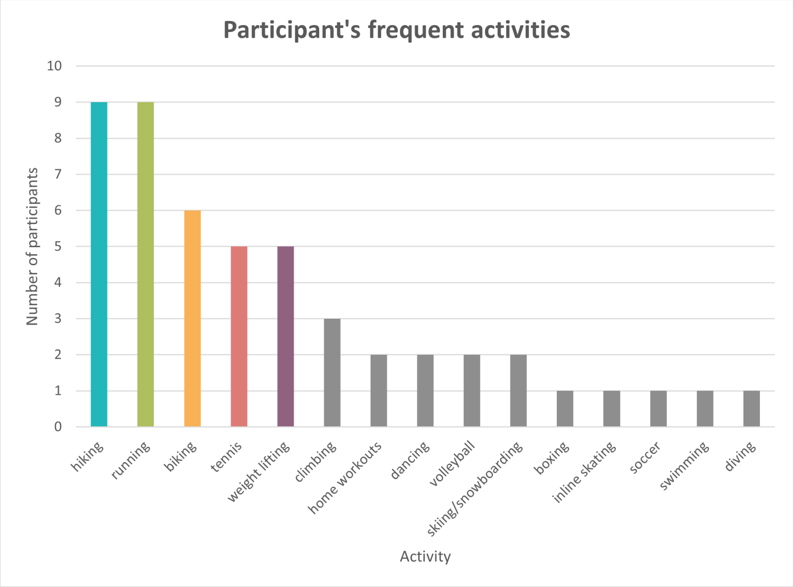


Figure 5.3.: Frequent activities

5.1.3. Smartphone Use for Health Related Purposes

Additionally, information about the use of health related applications is gathered. More than half of the participants are using apps for health related purposes. 9 out of the 20 participants are not using such applications as shown in Figure 5.4.

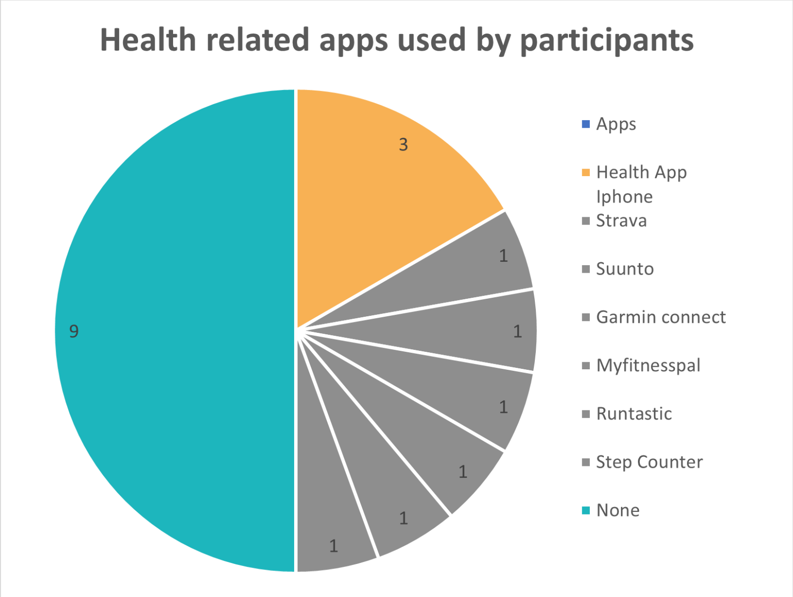


Figure 5.4.: Health apps

5.1.4. Perceived Usability

Furthermore, the PSSUQ is used to measures the perceived satisfaction of participants when using the ***aktivtalk*** application.

For the System Usefulness (SYSUSE) (Figure 5.5) and the Interface Quality (INTERQUAL) (Figure 5.6) the mean values and the standard deviations (SYSUSE: M=1.28, SD=1.29; INTERQUAL: M=1.23, SD=0.642) are low which suggests that the majority of the participants strongly agree with the statements for SYSUSE and INTERQUAL. This indicates a high System Usefulnes and Interface Quality.

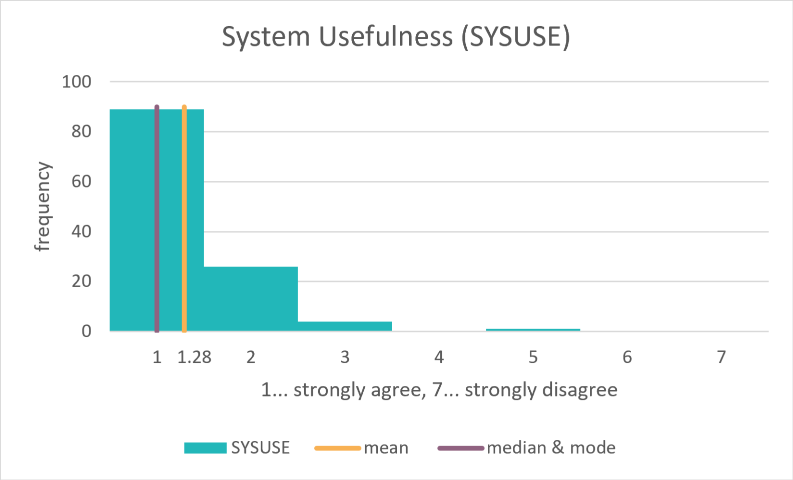


Figure 5.5.: System Usefulness (SYSUSE)

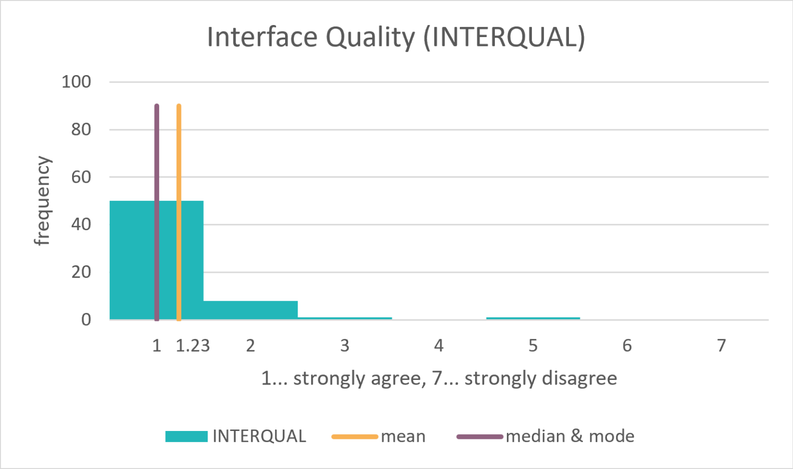


Figure 5.6.: Interface Quality (INTERQUAL)

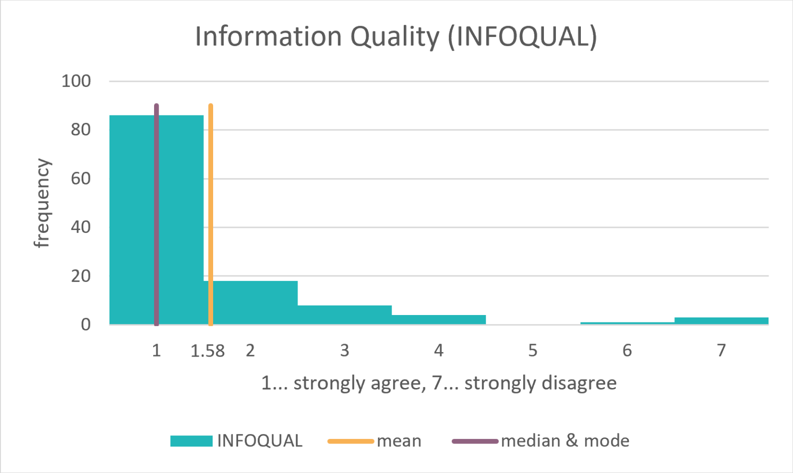


Figure 5.7.: Information Quality (INFOQUAL)

The Information Quality (INFOQUAL) (Figure 5.7) has a mean value of 1.58 and a higher standard deviation of 1.22 which indicates more variability in responses compared to INTERQUAL.

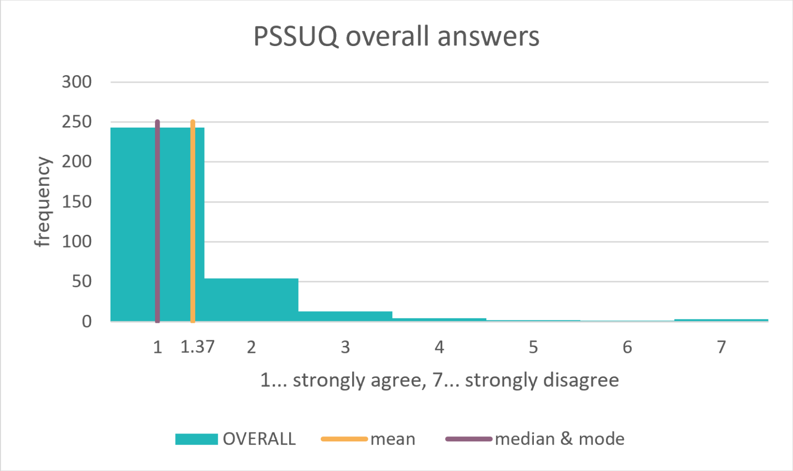


Figure 5.8.: PSSUQ overall answers

In the overall answers of the PSSUQ the results indicate that participants strongly agree with the system's usefulness, information and interface quality, as illustrated in Figure 5.8. The lower the mean the better the performance and satisfaction. It can be concluded that the ***aktivtalk*** application achieved a high satisfaction across the participants which addresses RQ3. The stacked bar chart below (Figure 5.9) shows the answers of the single questions:

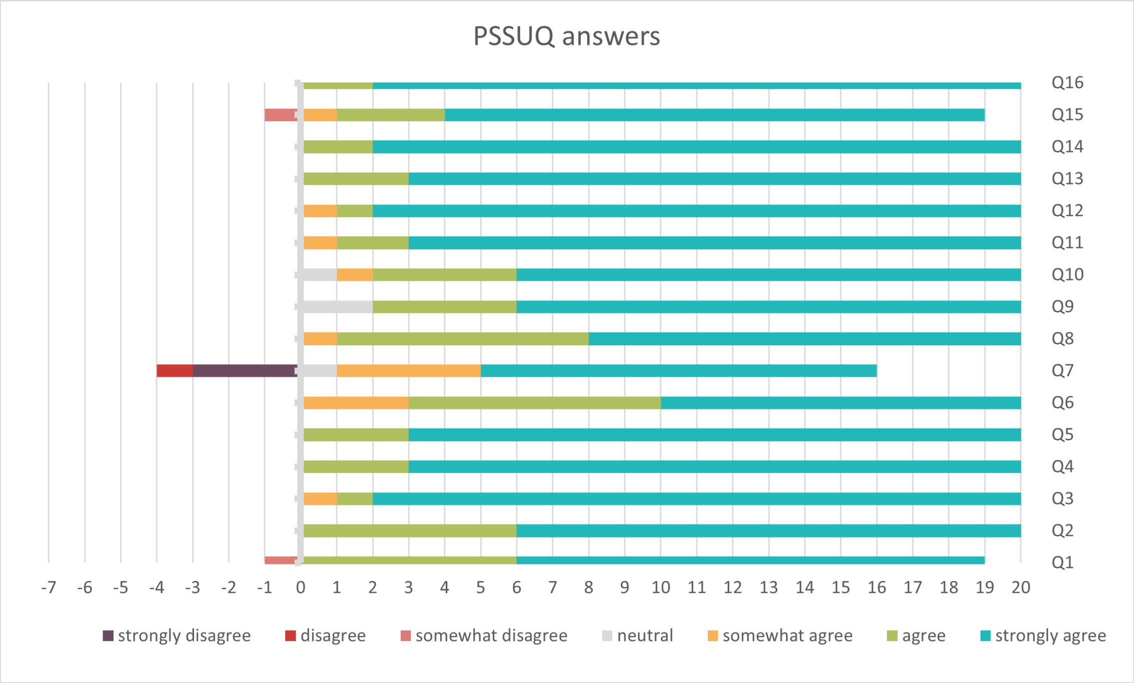


Figure 5.9.: PSSUQ answers

The ease of use of the different methods was investigated on a scale from 1 (very easy) to 5 (very hard). For the ***aktivtalk*** application with the YNNS assessment the mean value is 1.55 (see Figure 5.10), which means that the participants found it very easy to moderately easy to assess themselves with the YNNS assessment in the application. The range (2) and standard deviation (0.74) suggest some variability in the responses, but the majority of the participants still found it very easy.

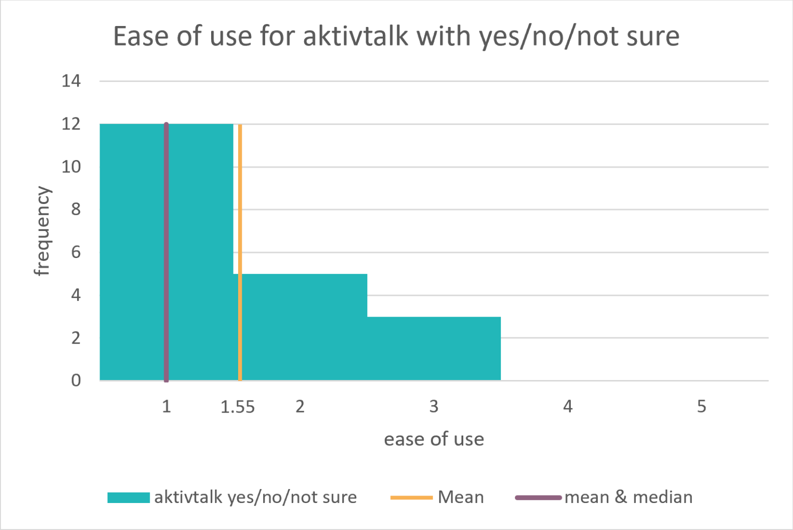


Figure 5.10.: Ease of use YNNS

The ease of use for the ***aktivtalk*** application with the BORG scale assessment has a mean of 1.65 (Figure 5.11) which indicates that the participants also found the assessment very easy to moderately easy. The range (3) and standard deviation (0.85) is slightly higher than for the YNNS assessment.

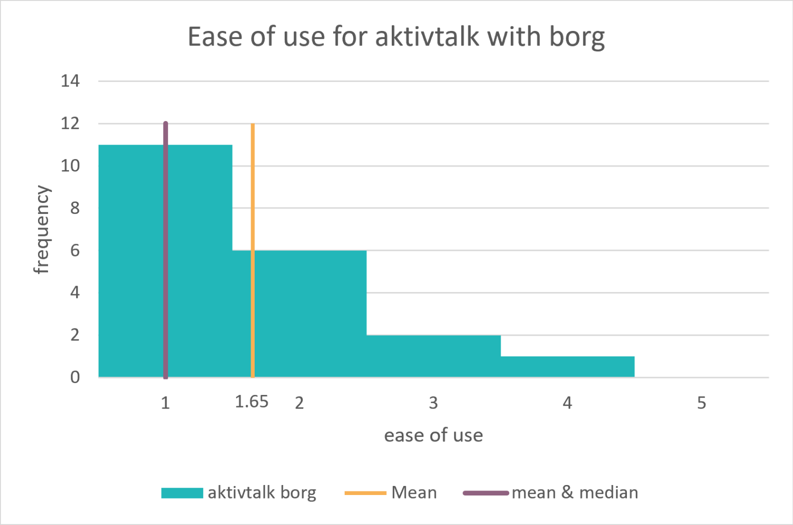


Figure 5.11.: Ease of use BORG

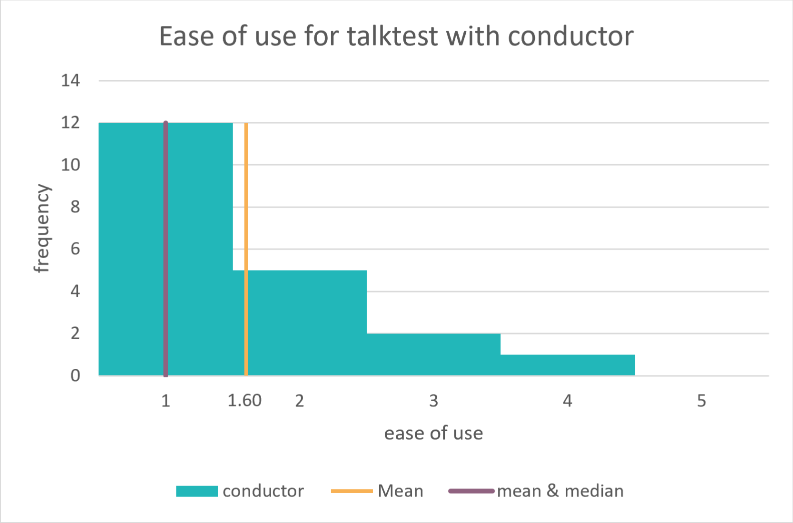


Figure 5.12.: Ease of use conductor

For the self assessment with the conductor the mean value is 1.60 (Figure 5.12) which is slightly lower than for the BORG scale assessment. The range (4) and standard deviation (0.86) is the highest across the different methods, however, the majority also found it very easy to assess themselves with the conductor.

There are only slight variations in mean, range and standard deviation across the three methods. In general, all methods are considered easy to use by the participants.

5.1.5. Interviews

The data from the interviews is extracted by re-listening and writing down the key points of what the participants stated. The key points are then summarized into categories, in order to subdivide the statements. The participants are asked about what they like most in the ***aktivtalk*** application. 13 answered that the app is easy to use and understand and 12 stated that the user interface is appealing. 3 mentioned the countdown, 2 minimized clicks and another 2 usability in the context of the best aspect (Figure 5.13).

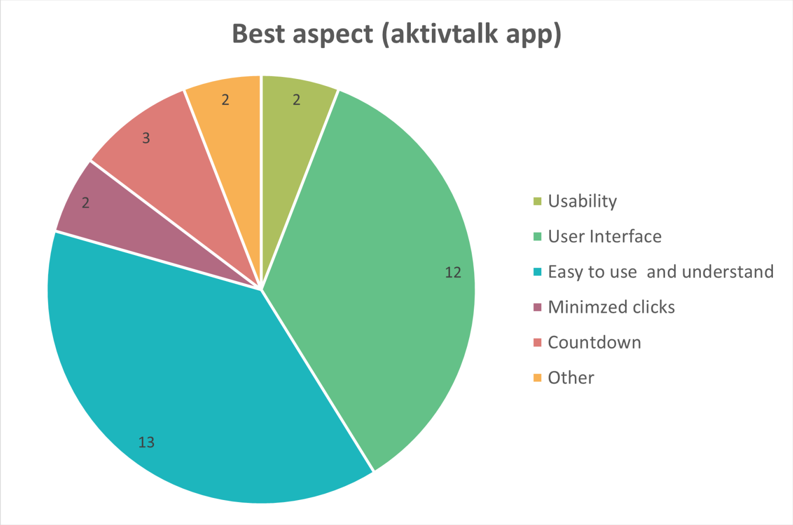


Figure 5.13.: Best aspect

To the question what needs most improvement in the application 5 participants answered that it is hard to choose on the BORG scale because the slider is too small and with a higher level of exhaustion the cognitive skills become more inconspicuous. 4 stated that the paragraphs could have a bigger variety, adaptable font-sizes and lower complexity. Furthermore, another 4 stated that the button and font size could be increased in general since it becomes harder to read when the intensity level increases (Figure 5.14).

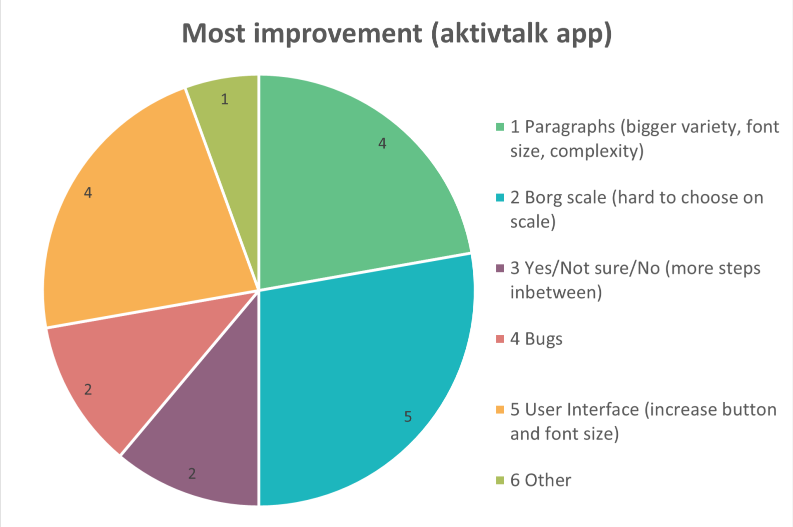


Figure 5.14.: Most improvement

The participants are also asked whether they trust the different methods or not. None of them answered that they do not trust any method at all. They assumed that the methods are trustful, however, they compared the methods against each other. The results can be seen down below in Figure 5.15 and Figure 5.16.

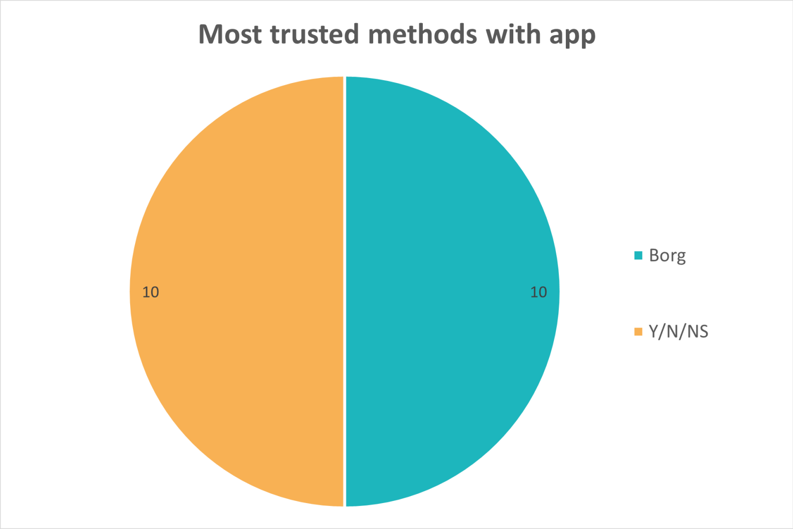


Figure 5.15.: Most trusted methods with app

An equal amount of participants answered that they trust the BORG and YNNS assessment more than the conductor guided assessment.

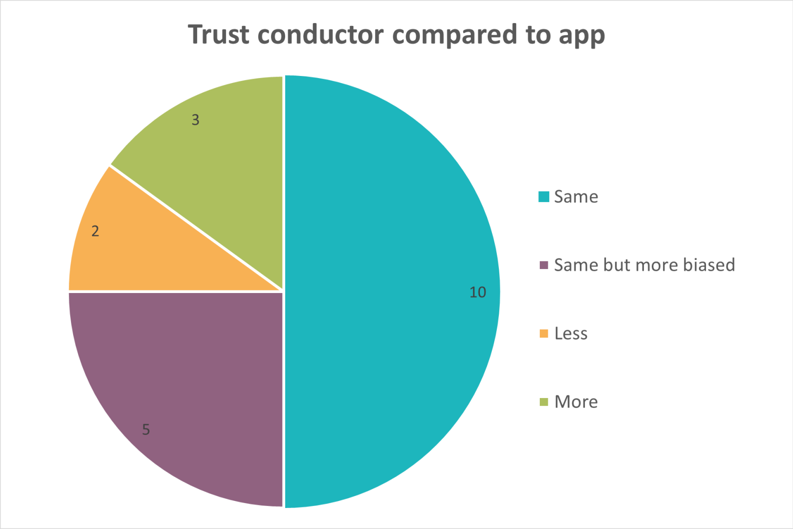


Figure 5.16.: Trust conductor compared to app

The participants compared the trust of the method with the conductor against the methods with the ***aktivtalk*** application. In total 15 stated that they have the same trust in the method with the conductor, however, 5 of them feel more biased with the conductor compared to the methods with the application. 3 participants have more trust and 2 participants have less trust during the conductor guided self-assessment (Figure 5.16).

5.1.6. Method Ranking

The participants further ranked the methods according to their most to least favorite method (see Figure 5.17). For each participant the methods receive points according to their ranking. The first receives 2, the second receives 1 and the third receives 0 points. The points received from each participant are then summed up and divided by the number of participants.

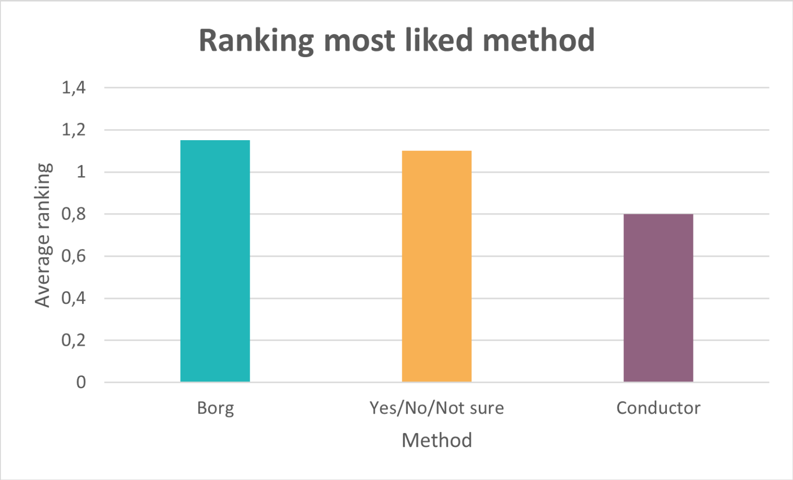


Figure 5.17.: Ranking most liked method

From the above Figure 5.17 it can be observed that the BORG method achieves the highest average ranking, closely followed by the YNNS approach. The method with the conductor is the least liked method across the three methods.

5.1.7. Observations and Adaptions

It is observed that in the beginning some participants had trouble when assessing themselves on a scale from 6 to 20 or answering with yes, not sure or no to the question if talking was still comfortable. By explaining the participants the options to answer from and the exercise intensity zones, they got more comfortable when answering. This effect could also be mitigated by introducing help buttons and an introduction page in the ***aktivtalk*** application which could affect the participants' confidence when answering in a positive manner.

Furthermore, it appears that the participants tend to give more biased answers when Method C is used. This effect is also observed by 5 out of 20 participants in the interviews (see Figure 5.16). However, from the average overlaps in Section 5.2 it can be derived that the participants answer in a more confident way when the conductor is present, since the values and standard deviations are smaller compared to other methods.

In Subsection 3.3.8 it is stated that the participants are free to choose their intensity level. However, participants tend to stay in a lower intensity level. Since samples of high intensity are required in the same way as samples of lower intensities, the intensity was adapted and increased by the conductor in consultation with the participants. The conductor tried to get an intensity peak within each method, however, this was not possible for every participant due to their fitness level and previous exhaustion.

Moreover, the initial plan was to use a testing interval of 2 minutes. The following calculation shows the total number of samples:

$$P \text{... Nr. of participants}$$

$$TI \text{... Testing interval}$$

$$WT \text{... Workout time}$$

TT... Talk Test and Assessment Time

SP... Samples per Person

ST... Samples total

$$\frac{WT * 60}{TI * 60 + WT} = SP * P = ST$$

$$\frac{35 * 60}{2 * 60 + 30} = 14 * 20 = 280 \text{ samples}$$

The calculation above shows that with a 2 minute testing interval the total amount of samples would only be 280, assuming that every sample can be used. To generate additional samples, a testing interval of 1 minute is implemented. The frequency of testing increases when the participant approves and when the intensity is higher. This adjustment in the protocol is made based on observations from the initial participants, revealing that maintaining participants at high intensity for shorter durations while testing more frequently is more convenient. This enables generating a higher number of samples with high intensity even for participants with lower physical fitness levels who can not sustain high intensity for extended durations.

This strategy was not initially implemented because there was uncertainty about whether frequent testing would be too stressful for participants. However, this proved not to be the case.

5.2. Audio and Pulse Analysis

When analyzing the synchronized data from the participants, overlaps within the pulse ranges corresponding to light, medium, and high-intensity exercise categories are observed. The underlying reason for these overlaps may arise from participants' potential discomfort in self-assessment. However, the different answers for the YNNS and BORG method are explained to them in detail before the workout.

Due to the observed overlaps between self-assessed answers and corresponding pulse ranges, the target heart rate and intensity zones for each participant based on their age are calculated. A more precise approach for calculating the heart rate intensity zones would require the determination of each participant's maximum heart rate. However, this level of granularity falls outside the scope of the thesis and therefore, the heart rate intensity zones, relative to age, are determined as follows [25]:

1. Max Heart Rate (MHR) = 220 - Age
2. Target Heart Rate (THR) = MHR * %Intensity

Target Zone	% Intensity
Maximum	90% - 100%
Hard	80% - 90%
Moderate	70% - 80%
Light	60% - 70%
Very Light	50% - 60%

Table 5.1.: Intensity zones according to maximum heart rate [25]

In the above Table 5.1 the target zones can be seen. Heart rates exceeding the upper limit of the hard target zone (maximum) are categorized as high intensity, while heart rates falling below the lower limit of the light target zone (very light) are designated as light intensity.

After computing the according target zone for every sample, the heart rate labels for 2 and 3 classes are added to the dataset as well. A small snipped of the dataset can be seen below in Table 5.2:

filename	label_3c	label_2c	avg_pulse	label_pulse_3c	label_pulse_2c
P13c3-stageYes.wav	2	1	148	1	1
P13c11-stageNotSure.wav	1	1	180	0	0
P13y32-stageNo.wav	0	0	190	0	0

Table 5.2.: Dataset snippet

In Table 5.2 it can be clearly observed that the labels for the average pulse ("label_pulse_3c" and "label_pulse_2c") differ from the self-predicted labels ("label_3c" and "label_2c"). The reason for this disparity may arise from the lack of self-assessment confidence, or the fact that the target pulse according to the age is not representative for every participant. These significant differences in the self-assessed intensity zones and pulse intensity zones are not applicable to all participants and samples.

Moreover, the minimum and maximum pulse for all methods for each participant are used in order to calculate the overlaps for the light-medium and medium-high intensity zone. The average overlaps and the standard deviations can be seen in Table 5.3 below:

Overlaps for self-assessed labels and pulse in bpm						
	y/ns/n		**BORG**		**conductor**	
	mean	SD	mean	SD	mean	SD
low-medium	8.22	4.80	8.01	5.67	7.32	5.64
medium-high	7.95	5.72	8.08	7.44	5.06	3.47

Table 5.3.: Mean and standard deviation of overlaps of self-assessed labels and pulse across methods

It is evident from Table 5.3 above that the mean overlaps for the self-assessed pulse with the conductor guided method are the smallest (overlaps conductor: low-medium M=7.32; medium-high M=5.06). Furthermore, the standard deviation is smaller compared to the BORG scale assessment which indicates that the pulse values are less spread out (overlaps conductor: low-medium SD=5.64; medium-high SD=3.47). This could suggest that the participants are more confident in choosing their right intensity zone when the conductor is guiding the workout. For the YNNS assessment, the average of the low-medium overlap is relatively high (overlaps YNNS: low-medium M=8.22, SD=4.80). The reason for this higher value could be that the participants are unsure about whether they are still in the low or medium intensity zone when assessing themselves. For the medium-high overlaps, the values are smaller which could indicate more confidence in assessing with YNNS when the intensity is higher (overlaps YNNS: medium-high M=7.95, SD=5.72). For the BORG scale assessment, the average overlap and standard deviation for medium-high are slightly higher compared to the values from the YNNS assessment (overlaps BORG: medium-high M=8.08, SD=7.44).

Table 5.4 includes descriptive statistics of the pulse for all methods and intensity zones:

		min	**mean**	**max**	**SD**
y/ns/n	Light	75.13	130.21	166.75	20.39
	Medium	116.60	151.71	185.53	15.66
	High	126.93	165.05	190.00	13.98
BORG	Light	85.25	126.76	162.88	18.19
	Medium	111.53	147.15	177.60	15.79
	High	135.53	162.51	179.81	12.16
c	Light	83.20	131.94	176.19	15.90
	Medium	121.87	149.83	177.00	11.17
	High	127.60	162.14	180.73	12.44

Table 5.4.: Descriptive statistics of pulse for all methods and intensity zones

The above Table 5.4 also shows a higher standard deviation for the intensity zones of the YNNS assessment (YNNS: light SD=20.39, medium SD=15.66, high SD=13.98). The standard deviations are the highest across all methods for the light and high intensity zones of the YNNS assessment. For the medium intensity zone of the YNNS assessment, the standard deviation is close to the highest value which belongs to the BORG scale assessment (BORG: medium SD=15.79). This could also indicate that the participants are less confident in assessing themselves when using the YNNS method. The conductor guided method has the lowest standard deviations for the light and medium intensity zone (c: light SD=15.90, medium SD=11.17).

5.3. Model Evaluation

In theory, supersampling appeared to be a promising approach to generate more samples. However, this method did not result in performance improvement for the machine learning model. For this reason, supersampling was not used when doing the evaluation.

Since the dataset is imbalanced, various oversampling techniques were investigated in order to generate a balanced number of samples across the classes. However, the accuracy and loss did not improve by using the oversampling techniques SMOTE, ADASYN, SVMSMOTE and BorderlineSMOTE. Using SMOTEENN improved the accuracy and loss drastically. The reason for this could be that SMOTEENN was effective in reducing the number of noisy samples in the dataset, which led to improved accuracy.

In this study the classification of exercise intensity through speech involved the application of a machine learning approach, specifically employing a Sequential Neural Network. Other machine learning approaches where investigated for the classification of voice samples. For example, a Convolutional Neural Network (CNN) was utilized, however, its predictive performance was observed to be suboptimal compared to the Sequential Neural Network. Despite the complexity of the classification task, commonly suited for Long Short-Term Memory (LSTM) networks due to their ability to capture long-range dependencies, the initial LSTM model surprisingly performed worse on the dataset compared to the Sequential Neural Network. The limited performance of both the CNN and LSTM led to the choice of the Sequential Neural Network as the preferred architecture for classifying exercise intensity.

The accuracy and loss metrics are determined by averaging the scores of the 5-fold cross-validation. Moreover, the model is executed five times to observe additional iterations.

5.3.1. Performance of 3 Classes

In Figure 5.18 and Figure 5.19 the accuracy and loss of the model for 3 classes can be seen. The three classes are used in order to distinguish between the light, medium and high intensity zones.

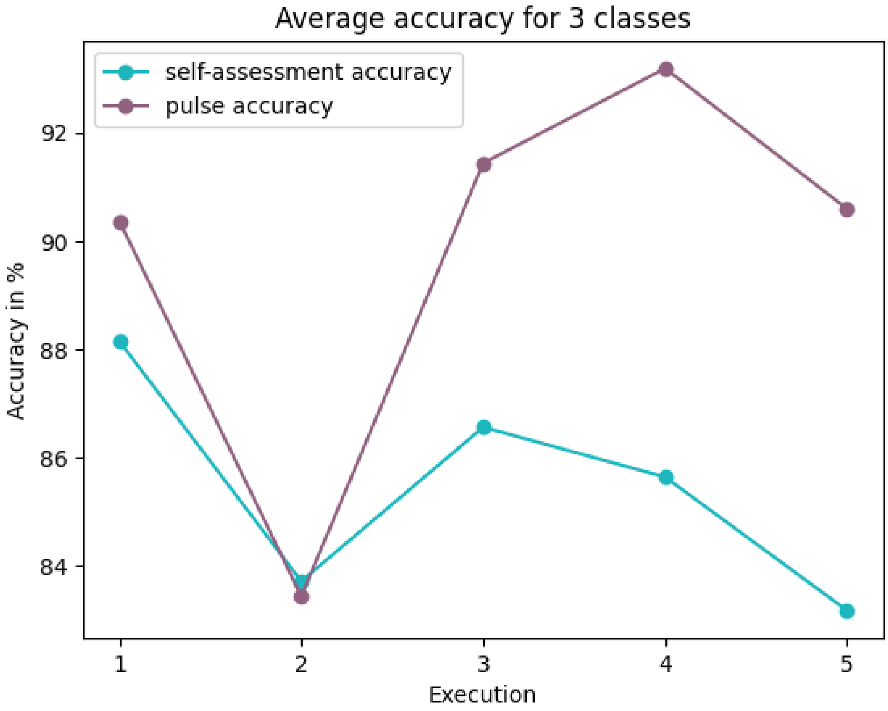


Figure 5.18.: Accuracy for 3 classes

- Self-assessment accuracy: The model achieved an average accuracy ranging from approximately 83.18% to 88.15%.
- Pulse accuracy: The model achieved an average accuracy ranging from approximately 83.44% to 93.19%. The accuracy values are generally higher than those for the self-assessment.

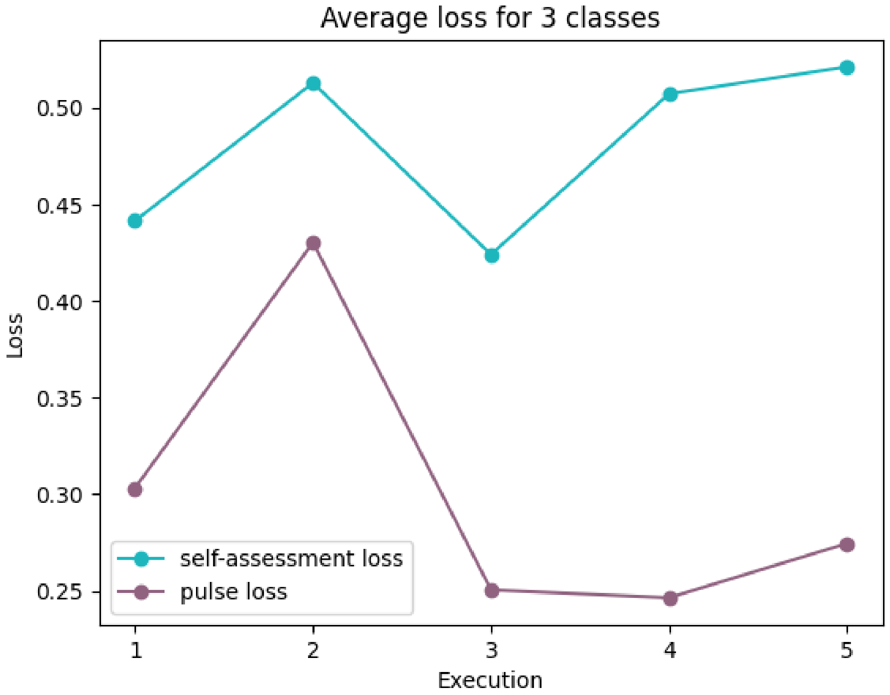


Figure 5.19.: Loss for 3 classes

- Self-assessment loss: The average loss values range from about 0.424 to 0.521.
- Pulse loss: The average loss values range from about 0.250 to 0.431. The loss values are generally lower than those for the self-assessment.

From Figures 5.18 and 5.19 it can be observed that the model performs better on the dataset with the labels regarding the pulse and the age target zone. The accuracy for the pulse data is higher and the loss lower, which indicates better performance.

5.3.2. Performance of 2 Classes

In Figure 5.20 and Figure 5.21 the accuracy and loss of the model for 2 classes can be seen. The two classes are used in order to only distinguish between the light/medium and high intensity zones. This is sufficient for the automated exercise intensity estimation of patients with cardiovascular diseases, since it is important for the patients to not work out in the high intensity zone.

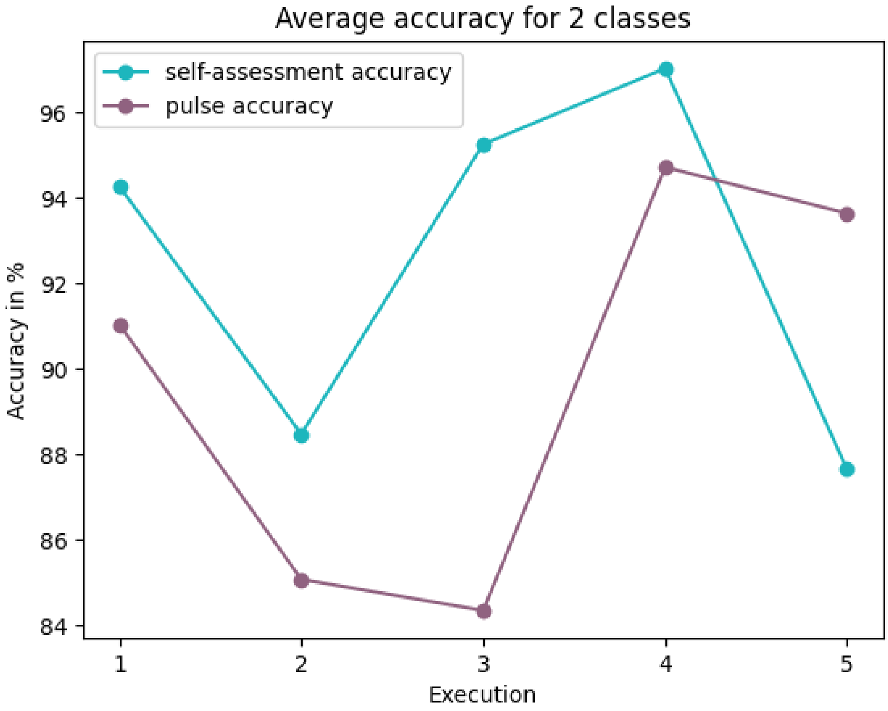


Figure 5.20.: Accuracy for 2 classes

- Self-assessment accuracy: The model achieved an average accuracy ranging from 87.64% to 97.02%. This indicates that the model's performance on the self-assessment is quite good, with high accuracy values.

- Pulse accuracy: The model achieved an average accuracy ranging from about 84.34% to 94.71% across the 5-fold cross-validation. The accuracy is slightly lower than that of the self-assessment.

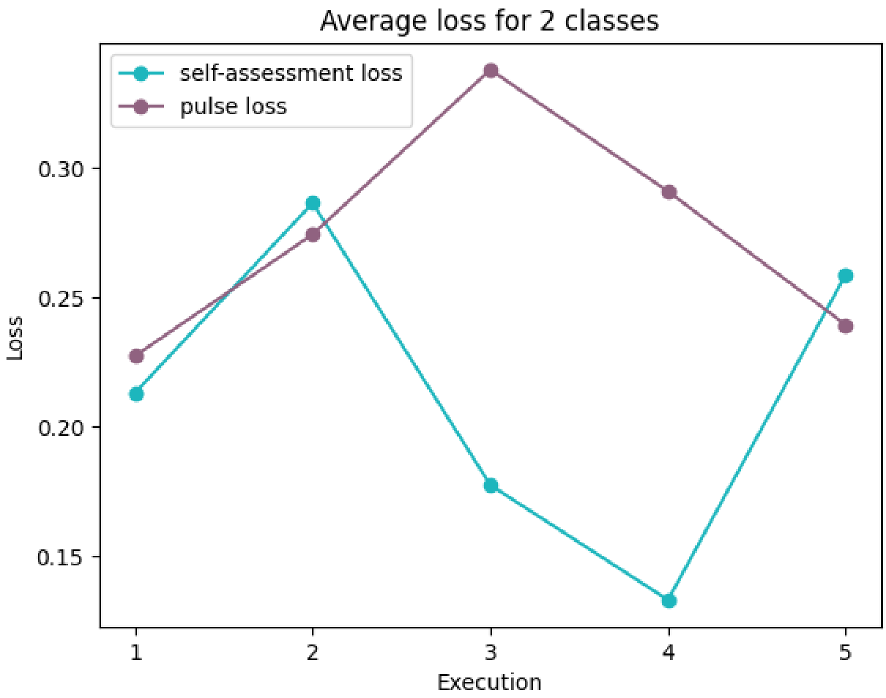


Figure 5.21.: Loss for 2 classes

- Self-assessment loss: The average loss values are relatively low, ranging from about 0.133 to 0.287.
- Pulse loss: The average loss values range from approximately 0.228 to 0.338. These values are higher compared to the self-assessment loss, indicating that the model's fit to the pulse loss is not as tight as it is to the self-assessment loss.

From Figures 5.20 and 5.21 it can be observed that the performance of the model with 2 classes and the self-assessment performs better than the model with 2 classes and the pulse target zone labels. The average accuracy is higher and the average loss is lower, which indicates a better performance.

6. Discussion

The reviewed studies in Section 4.1.2 explored the medical background of the Talk Test using consistent testing procedures on a cycle ergometer, with variations in warm-up and load increments. Despite different terminology for activity modes, a strong correlation exists due to similar protocols. Participants had diverse diagnoses, and study durations varied, with one lasting 8 weeks to assess changes in physical fitness. The studies used different measurement metrics and variations in power output influenced by factors like diagnosis, gender, and participant numbers were observed. In examining the relationship of the Talk Test with the Ventilatory Threshold, inconsistent findings were noted. However, the positive and equivocal stages of the Talk Test align with the prescribed intensity zones for patients. Overall, the research highlights the need for standardization and further exploration of the Talk Tests utility in assessing physical fitness, especially in a home-based setting.

The usability study of the ***aktivtalk*** application employed the PSSUQ to measure the perceived satisfaction. It revealed that the majority of participants strongly agreed with statements related to System Usefulness (SYSUSE) and Interf ace Quality (INTERQUAL), reflecting high levels of satisfaction in these aspects. However, Information Quality (INFOQUAL) showed more variability in responses. Overall, participants displayed a strong agreement with the application's usability, indicating a high level of satisfaction. Furthermore, the study assessed the ease of use of different methods, including YNNS assessments and BORG scale assessments. The mean values indicated that participants found all methods to be relatively easy to use, with minor variations in the range and standard deviation. This suggests that the participants perceived all methods as user-friendly, providing valuable insights into their usability and applicability.

Moreover, participants' responses highlighted what they liked most and areas in need of improvement in the ***aktivtalk*** application. Specifically, the ease of use, user interface appeal, and specific features were positively noted. Feedback regarding aspects requiring improvement, such as slider size and readability, was also gathered. Participants' trust in the assessment methods using the application was equal. However, one quarters of the participants stated that the either have more or less trust in the conductor guided method compared to the methods using the ***aktivtalk*** application. This might be because of the conductor's presence, which could give a wrong sense of trust.

Participants ranked the assessment methods according to their preferences, with the BORG method receiving the highest average ranking, followed closely by the YNNS approach. In contrast, the conductor-guided method received lower rankings overall. The reason for this could be that the ***aktivtalk*** application provides a sense of autonomy and self-directed engagement, allowing participants to navigate assessments at their own pace and convenience. The user interface might also enhance the overall experience by making the Talk Test more accessible without relying on external guidance.

The synchronization with the pulse showed that there were some overlaps of the pulse and the self-assessed intensity areas. The conductor guided assessment had the smallest number of average overlaps, which could indicate that the participants are more confident with assessing themselves when the conductor is present. However, this is not aligning with the results from the interviews about trust. Figure 5.16 shows that 5 participants have the same trust in the conductor guided method but feel more biased when answering. This discrepancy could arise from the optimism about using technology such as the ***aktivtalk*** application. This bias can influence subjective judgments and preferences even when objective data suggests a different reality.

From Figure 5.18 and Figure 5.20, it can be observed that the performance of the model does not undergo a substantial transformation when using 2 labels instead of 3 for the pulse. For example, the reason for this could be that the model has to be adapted for this kind of classification or there is not enough quality sufficient data. Furthermore, another interesting fact is that the model performance with the self-assessed intensity zones improves when using 2 instead of 3 labels. This might be the case due to minor inconsistencies and mislabeling in the samples for the classification with 3 labels. In this evaluation, it can be observed that using the heart rate target zones is not the most accurate way to compute the perceived intensity of the participants. However, the performance of the self-assessed intensity zones is close to the one using the pulse with 3 labels and even outperforms the pulse when using only 2 labels. The fact that the model performs better with a simplified classification system suggests that the self-assessed intensity zones, while not perfectly aligned with heart rate target zones, can offer a more robust and accurate representation of perceived intensity.

Addressing RQ4, this initial machine learning model is able to detect intensity zones with an accuracy up to 97.02%. However, to improve the performance of the model, the dataset has to be expanded and extensive work on feature engineering has to be done.

7. Limitations

7.1. Systematic Literature Review

Only one researcher has conducted the Systematic Literature Review, since it was done in the scope of a master thesis. Limitations such as potential bias in study selection, interpretation due to individual perspectives and the risk of overlooking relevant literature due to resource constraints and time limitations could arise. Additionally, the absence of peer review in the selection and synthesis process may impact the review.

7.2. Digital Talk Test

The execution possibilities of the ***aktivtalk*** application are limited as it is not publicly available for the end-user yet. Developed for prototyping purposes, the application needs to undergo extensive testing and detailed documentation before deployment.

7.3. User Study

It is important to note that participants may modify their behavior during the workout due to the presence of the conductor, introducing a potential bias in self-assessment. This could lead to participants either overestimating or underestimating their performance under these observed circumstances.

Furthermore, the audio quality could be improved by using better recording devices. This might be beneficial for artificial situations where the audio quality needs to be of high intensity. However, for real life scenarios the audio quality is not from professional recording devices and there might also be some noise in the background.

7.4. Data and Machine Learning

The dataset consists of 559 samples which is an insufficient amount of data regarding the complexity of the classification task. Moreover, the dataset is not representative of the entire population and therefore, the model might produce biased results. The limitation regarding the size of the dataset, arises from practical constraints within the context of this master thesis. Conducting extensive data collection involving a more significant number of samples requires resources, time, and logistical support that are beyond the scope of a master thesis. The complexity of the classification task further amplifies the challenge, as a more intricate task generally necessitates a larger dataset for robust model training.

8. Future Work

The Talk Test has been explored in past literature and its usefulness for patients with cardiovascular diseases has been investigated. The study of this thesis does not contain any patients with cardiovascular diseases. It is solely conducted on participants recruited from the circle of acquaintances, since the major goal is to collect audio samples for an initial machine learning model. In future work it is recommended to perform a study with cardiovascular patients, especially those taking beta-blockers. Furthermore, the participants should also be evenly distributed among their age groups.

In the context of the protocol and the actual study execution, it is advisable to establish a mechanism to facilitate the synchronization of the pulse with the audio samples in order to reduce the workload after the conduction. Furthermore, the maximum heart rate should be clinically measured in order to get a more precise estimation of the intensity zones during the workout. Furthermore, in previous literature the speech comfort level was observed by the researcher or clinician most of the time. This might be one of the reasons why the Talk Test is considered to be a valid tool for estimating the current intensity zone. Therefore, for future research and studies, the speech comfort level should also be verified by the study conductor in order to explore if there is a major difference between the self-assessment and the observation from the conductor. This could also help to detect and analyze a learning bias when using the Talk Test and especially the ***aktivtalk*** application.

In order to make sure the prototype of the ***aktivtalk*** application is production ready, the work described above, has to be done initially. Moreover, extensive testing, including more user tests, has to be done.

Additionally, the machine learning model has to be further developed to achieve better performance results. For example, more features could be extracted from the audio files including breathing rate, sentiment etc. This might improve the overall accuracy as well.

The initial idea was to use the machine learning model in the ***aktivtalk*** application, in order to enable an option to assess the current exercise intensity level without the need of the user to assess. This feature could improve the usability and user experience intrinsically, since the user is required to interact with the application less and there is no bias due to mislabeled audio samples by the users themselves. However, this requires a well performing machine learning model.

9. Conclusion

In conclusion, this master thesis has addressed the critical issues surrounding the assessment of exercise intensity in patients with cardiovascular diseases, with a specific focus on the development and evaluation of a Digital Talk Test, called ***aktivtalk***, for the Ludwig Boltzmann Institute for Digital Health and Prevention.

From the medical background of the Talk Test and the identified literature it can be concluded that the Talk Test is a reliable tool for assessing the exercise intensity of patients with various diagnoses including cardiovascular diseases. This background builds a baseline for the development of the ***aktivtalk*** which proved to be useful for conducting the Talk Test and collecting voice samples from participants. Furthermore, the user study showed that the application achieved high levels of user satisfaction, as indicated by user feedback and assessments of system usefulness, information quality, interface quality and trust.

The pulse overlaps show some inconsistencies with the confidence of the self-assessment across the methods. However, it is not clear where these inconsistencies arise from, since multiple factors could influence the results. The development of a machine learning model for detecting and classifying exercise intensity levels from audio files collected by the Digital Talk Test was explored. The evaluation of the model indicated the potential of this system to effectively categorize exercise intensity levels.

The findings of this master thesis suggest that the ***aktivtalk*** application has the potential to be a valuable tool in the assessment and management of exercise intensity for patients with cardiovascular diseases. This digital tool not only shows promise in terms of its reliability and usability but also hints at its potential for further development in the area of digital healthcare solutions. As such, the results presented in this thesis contribute to the growing knowledge in the field of cardiovascular rehabilitation and digital health technologies, offering a promising future for improving patient care.

A. Rapid Assessment of Physical Activity (RAPA)

	Yes	No
1. I rarely or never do any physical activities.	□	□
2. I do some light or moderate physical activities, but not every week.	□	□
3. I do some light physical activity every week.	□	□
4. I do moderate physical activities every week, but less than 30 minutes a day or 5 days a week.	□	□
5. I do vigorous physical activities every week, but less than 20 minutes a day or 3 days a week.	□	□
6. I do 30 minutes or more a day of moderate physical activities, 5 or more days a week.	□	□
7. I do 20 minutes or more a day of vigorous physical activities, 3 or more days a week.	□	□
8. I do activities to increase muscle strength, such as lifting weights or calisthenics, once a week or more.	□	□
9. I do activities to improve flexibility, such as stretching or yoga, once a week or more.	□	□

B. Post-Study System Usability Questionnaire (PSSUQ)

	Strongly Agree						Strongly Disagree
	1	2	3	4	5	6	7
1. Overall, I am satisfied with how easy it is to use this system.	□	□	□	□	□	□	□
2. It was simple to use this system.	□	□	□	□	□	□	□
3. I was able to complete the tasks and scenarios quickly using this system.	□	□	□	□	□	□	□
4. I felt comfortable using this system.	□	□	□	□	□	□	□
5. It was easy to learn to use this system.	□	□	□	□	□	□	□
6. I believe I could become productive quickly using this system.	□	□	□	□	□	□	□
7. The system gave error messages that clearly told me how to fix problems.	□	□	□	□	□	□	□
8. Whenever I made a mistake using the system, I could recover easily and quickly.	□	□	□	□	□	□	□
9. The information (such as online help, on-screen messages, and other documentation) provided with this system was clear.	□	□	□	□	□	□	□
10. It was easy to find the information I needed.	□	□	□	□	□	□	□
11. The information was effective in helping me complete the tasks and scenarios.	□	□	□	□	□	□	□
12. The organization of information on the system screens was clear.	□	□	□	□	□	□	□
13. The interface of this system was pleasant.	□	□	□	□	□	□	□
14. I liked using the interface of this system.	□	□	□	□	□	□	□
15. This system has all the functions and capabilities I expect it to have.	□	□	□	□	□	□	□
16. Overall, I am satisfied with this system.	□	□	□	□	□	□	□

C. Ethics Approval

LUDWIG
BOLTZMANN
GESELLSCHAFT

Ref: 006_2023

To whom it may concern,

this is to confirm that the project

A Digital Talk Test for Assessing Exercise Intensity of Patients with Cardiovascular Diseases

led by

Laura Geiger, BSc (Ludwig Boltzmann Institute for Digital Health and Prevention, Salzburg)

has been evaluated by the LBG Ethics Committee and received a favourable opinion.

Signed:

Date:
25.07.2023

Dr Steph Grohmann
Senior Program Manager Ethics
Ludwig Boltzmann Gesellschaft
Open Innovation in Science Center

D. Data Protection and Information Sheet for Participants

Zustimmung zur Datenerhebung und Datenverwertung

Studie: aktivtalk, digitaler Sprachtest für die Selbstbeinschätzung der Trainingsintensität

Studienleitung: Laura Geiger
Universität Innsbruck
Institut für Informatik
Technikerstraße, 6020 Innsbruck

Tel: +43 681 814 286 00
E-Mail: laura.geiger@dhp.lbg.ac.at

Kontakt LBI-DHP: Jan Smeddinck
Ludwig Boltzmann Institut für
digitale Gesundheit und Prävention
Lindhofstrasse 22, 5020 Salzburg

Tel: +43 (0) 5 7255 82711
E-Mail: jan.smeddinck@dhp.lbg.ac.at

Der Studie wird vom Ludwig Boltzmann Institut für digitale Gesundheit und Prävention (LBI) und der Universität Innsbruck durchgeführt. Die verantwortliche Studienleitung ist Laura Geiger. Ziel dieser Studie ist es, Sprachdateien während einer Trainingseinheit auf mehreren Intensitäts-Leveln zu sammeln. Zur Erhebung der Sprachdateien wird eine digitale Version des Talk Tests verwendet (mobile Applikation) und ein Fragebogen zur Benutzerfreundlichkeit inklusive kurzem Interview wird durchgeführt.

Sie erklären sich hiermit einverstanden, dass das Material (Fragebögen, Audio-Dateien und Interviews), das während der Studie erstellt und gesammelt wird, in weiterer Folge zur Auswertung bzw. Erarbeitung von wissenschaftlichen Untersuchungsergebnissen herangezogen werden darf. Das Material wird anonym ausgewertet und die auszuwertenden Rohdaten werden temporär (3 Monate) auf einem verschlüsselten Cloud-Speicher (Amazon Web Services) zwischengelagert. Dies dient der Erstellung eines maschinelles Lernmodells, welches automatisch die Trainingsintensität ohne Selbsteinschätzung erhebt. Zusätzlich wird das Material in pseudonymisierter Form dauerhaft verschlüsselt auf Datenträgern am LBI-DHP gespeichert.

Zusätzlich gibt es die Option zur Einverständnis der möglichen Veröffentlichung der aufgenommenen Audiodateien ohne Nennung des Namens. Die Zustimmung zur Veröffentlichung verbessert die allgemeine Reproduzierbarkeit der Studie und steigert den Beitrag und Wert für die Forschung. Wenn diese Option nicht gewählt wird, dann können die Daten nur für interne Trainingszwecke verwendet werden und werden hiermit nicht veröffentlicht.

Selektives anonymisiertes Rohmaterial bildet die Grundlage für Präsentationen, Beiträge und wissenschaftliche Publikationen bzw. kann zur beispielhaften Darstellung der wissenschaftlichen Tätigkeitsfelder beteiligter Projektmitglieder des LBI's verwendet werden. Dabei wird das Material immer zu ausreichendem Ausmaß anonymisiert, sodass es für Außenstehende keinesfalls möglich ist, Rückschlüsse zu Ihrer Identität zu ziehen. Zusätzlich werden diese Materialien und Daten am LBI-DHP dauerhaft verschlüsselt gespeichert.

Mit der Unterfertigung geben Sie Ihre unwiderrufliche Zustimmung, dass die im Zusammenhang mit dieser Studie gesammelten Daten und Materialien für den oben genannten Zweck verwendet werden dürfen und Sie daraus keinerlei Ansprüche gegen die Projekt- und Studienleitung oder andere beteiligte Partner ableiten können. Sie verpflichten sich des Weiteren dazu, sämtliche Informationen zu dieser Studie vertraulich zu behandeln und nicht an Dritte weiterzugeben.

Sie bestätigen, dass Sie an dieser Studie freiwillig teilnehmen. Zusätzlich bestätigen Sie, dass Sie über das Projekt und die Studie ausreichend informiert wurden, Ihre Fragen zu Ihrer Zufriedenheit beantwortet wurden und Sie jederzeit die Möglichkeit haben, von der Teilnahme zurückzutreten.

□ Veröffentlichung Audiodateien ohne Nennung des Namens

Name in Blockbuchstaben: ______________________________

Ort und Datum: ______________________________

Unterschrift: ______________________________

Allgemeine Informationen zur Studie und zur aktivtalk App

- **Ziel**
  - Erhebung der eigenen Trainingsintensität: Die aktivtalk App soll den User durch Sprachtests dabei unterstützen, die derzeitige Trainingsintensität zu erläutern und bei zu hoher Belastung, zur Verminderung der Intensität aufgefordert zu werden.
  - Verbesserung der Selbsteinschätzung: Durch das regelmäßige Erheben der eigenen Trainingsintensität soll die Selbsteinschätzung gestärkt werden und der User wird selbstbewusster beim trainieren ohne medizinischer Überwachung.

- **Ablauf**
  - Expliziter Nutzungswille: Wenn sich die Person explizit zur Aufnahme von Sprachdateien während einer Trainingsintensität auf einem Ergometer bereit erklärt, wird die Teilnahme gestattet.
  - Einführung aktivtalk App und Ergometer: Die teilnehmende Person wird am Tag der Studie in die aktivtalk App eingeführt und der Ergometer wird entsprechend angepasst.
  - Vorabfragebogen: Demografische Daten (Alter, Geschlecht), körperliche Aktivitäten und Häufigkeit, Smartphone-Nutzung (digitale Gesundheits-Apps) werden vor der Trainingseinheit erhoben.
  - Trainingseinheit: Eine 30-minütige Trainingseinheit inklusive regelmäßiger Sprachtests mit Selbsteinschätzung dient zur Erhebung von Sprachdateien.
  - Nachbefragung und Interview: Nach der Trainingseinheit wird ein Fragebogen zur Benutzerfreundlichkeit herangezogen.

- **Datenschutz und Datenverwertung**
  - Vertraulichkeit: Alle Information werden vertraulich behandelt.
  - Sichere Datenspeicherung: Die personenbezogenen Daten werden pseudonymisiert und auf sicheren Servern gespeichert (DSGVO konform). Das Material wird anonym ausgewertet und die auszuwertenden Rohdaten werden temporär (3 Monate) auf einem verschlüsselten Cloud-Speicher (Amazon Web Services) zwischengelagert.
  - Zweck der Datenverwertung: Die aufgenommenen Sprachdateien dienen zur Erstellung eines initialem Machine-Learning-Modells für die automatische Erläuterung der Trainingsintensität.

- **Freiwillige Teilnahme und Rücktrittsfreiheit**
  - Freiwilligkeit: Ihre Wahl die aktivtalk App zu verwenden ist voll und ganz freiwillig.
  - Beendigung jederzeit: Sie können die Teilnahme an der Studie jederzeit beenden, was die Löschung sämtlicher Daten von Ihnen beinhaltet.

- **Risiken**
  - Überanstrengung: Während des Trainings kann es zu Überanstrengung kommen. Aus diesem Grund wird darauf hingewiesen das Training bei Schwindelgefühl, Übelkeit oder sonstigen Beschwerden unverzüglich zu beenden.

- **Fragen**
  - Jederzeit Kontaktaufnahme: Bei Fragen kann jederzeit Kontakt aufgenommen werden (laura.geiger@dhp.lbg.ac.at).

Bibliography

[1] Aabo, M. R., Ragle, A.-M., Østergren, P. B. and Vinther, A. [2021], 'Reliability of graded cycling test with talk test and 30-s chair-stand test in men with prostate cancer on androgen deprivation therapy', *Supportive Care in Cancer* **29**, 4249–4256.

[2] Blair, S. N., Kohl, H. W., Barlow, C. E., Paffenbarger, R. S., Gibbons, L. W. and Macera, C. A. [1995], 'Changes in physical fitness and all-cause mortality: a prospective study of healthy and unhealthy men', *Jama* **273**(14), 1093–1098.

[3] Bok, D., Rakovac, M. and Foster, C. [2022], 'An examination and critique of subjective methods to determine exercise intensity: the talk test, feeling scale, and rating of perceived exertion', *Sports Medicine* **52**(9), 2085–2109.

[4] *Build apps for any screen* [n.d.], `https://flutter.dev/`. Accessed: 2023-10-23.

[5] *Cardiovascular diseases* [n.d.], `https://www.who.int/health-topics/cardiovascular-diseases`. Accessed: 2023-09-18.

[6] *Dart overview* [n.d.], `https://dart.dev/overview`. Accessed: 2023-10-23.

[7] Díaz-Buschmann, I., Jaureguizar, K. V., Calero, M. J. and Aquino, R. S. [2014], 'Programming exercise intensity in patients on beta-blocker treatment: the importance of choosing an appropriate method', *European journal of preventive cardiology* **21**(12), 1474–1480.

[8] Grimm, P. [2010], 'Social desirability bias', *Wiley international encyclopedia of marketing* .

[9] *How to Create Understand Mel-Spectrograms* [n.d.], `https://importchris.medium.com/how-to-create-understand-mel-spectrograms-ff7634991056`. Accessed: 2023-10-27.

[10] *Journal Rankings on Medicine* [n.d.], `https://www.scimagojr.com/journalrank.php?category=2701`. Accessed: 2023-09-18.

[11] Krawcyk, R. S., Vinther, A., Petersen, N. C. and Kruuse, C. [2017], '"graded cycling test with talk test" is a reliable test to monitor cardiovascular fitness in patients with minor stroke', *Journal of Stroke and Cerebrovascular Diseases* **26**(3), 494–499.

[12] Kumar, D. [2021], Designing and Evaluating Mobile Health Technology for Ambulatory Monitoring and Diagnosis of Heart Arrhythmias, Phd thesis, DTU Health Technology.

[13] Kumar, D., Maharjan, R., Maxhuni, A., Dominguez, H., Frølich, A. and Bardram, J. E. [2022], 'mcardia: a context-aware ecg collection system for ambulatory arrhythmia screening', *ACM Transactions on Computing for Healthcare (HEALTH)* **3**(2), 1–28.

[14] Nielsen, S. G., Buus, L., Hage, T., Olsen, H., Walsøe, M. and Vinther, A. [2014], 'The graded cycling test combined with the talk test is reliable for patients with ischemic heart disease', *Journal of cardiopulmonary rehabilitation and prevention* **34**(4), 276–280.

[15] Nielsen, S. G. and Vinther, A. [2016], 'Graded cycling test combined with the talk test is responsive in cardiac rehabilitation', *Journal of Cardiopulmonary Rehabilitation and Prevention* **36**(5), 368–374.

[16] Orizola-Cáceres, I., Cerda-Kohler, H., Burgos-Jara, C., Meneses-Valdes, R., Gutierrez-Pino, R. and Sepúlveda, C. [2021], 'Modified talk test: A randomized

cross-over trial investigating the comparative utility of two “talk tests” for determining aerobic training zones in overweight and obese patients’, *Sports Medicine-Open* **7**, 1–8.

[17] *PSSUQ (Post-Study System Usability Questionnaire)* [n.d.], `https://uiuxtrend.com/pssuq-post-study-system-usability-questionnaire/`. Accessed: 2023-10-30.

[18] *Rapid Assessment of Physical Activity (RAPA)* [n.d.], `https://depts.washington.edu/hprc/programs-tools/tools-guides/rapa/`. Accessed: 2023-10-30.

[19] Rodríguez-Marroyo, J. A., Villa, J. G., Pernía, R. and Foster, C. [2017], ‘Decrement in professional cyclists’ performance after a grand tour’, *International journal of sports physiology and performance* **12**(10), 1348–1355.

[20] Sedgwick, P. [2012], ‘The hawthorne effect’, *Bmj* **344**.

[21] *Shotcut is a free, open source, cross-platform video editor.* [n.d.], `https://shotcut.org/`. Accessed: 2023-10-23.

[22] *SMOTEENN* [n.d.], `https://imbalanced-learn.org/stable/references/generated/imblearn.combine.SMOTEENN.html`. Accessed: 2023-11-06.

[23] Sørensen, L., Larsen, K. S. R. and Petersen, A. K. [2020], ‘Validity of the talk test as a method to estimate ventilatory threshold and guide exercise intensity in cardiac patients’, *Journal of Cardiopulmonary Rehabilitation and Prevention* **40**(5), 330–334.

[24] *speech_to_text* [n.d.], `https://pub.dev/packages/speech_to_text`. Accessed: 2023-10-23.

[25] *Target Heart Rate Calculator* [n.d.], `https://www.calculatorsoup.com/calculators/health/target-heart-rate-zone-calculator.php`. Accessed: 2023-10-31.

[26] *The Mel Scale* [n.d.], `http://musicweb.ucsd.edu/~trsmyth/pitch2/Mel_Scale.html`. Accessed: 2023-10-27.

[27] *What is the Difference Between VT1, VT2 and VO2 max?* [n.d.], `https://www.acefitness.org/fitness-certifications/ace-answers/exam-preparation-blog/3139/what-is-the-difference-between-vt1-vt2-and-vo2-max/`. Accessed: 2023-09-18.

[28] Wonisch, M., Hofmann, P., Fruhwald, F. M., Kraxner, W., Hödl, R., Pokan, R. and Klein, W. [2003], 'Influence of beta-blocker use on percentage of target heart rate exercise prescription', *European Journal of Preventive Cardiology* **10**(4), 296–301.