Measuring pleasantness of bicyclists through heart rate variability over self-reports

MATHIJS TOBÉ, University of Twente, The Netherlands

Ultra-short-term (UST) heart-rate-variability (HRV) analysis in non-static conditions, such as bicycling, poses significant challenges. Most HRV analysis is traditionally conducted in a resting state, calling for an examination of UST HRV metrics applicable to our specific setting. We are trying to estimate the axis of valence in the circumplex model of affect, which relates to pleasantness. Consequently, certain HRV metrics may exhibit unreliability or insignificance in this estimation. Considering software and hardware limitations, we focus on time-domain HRV metrics that can be derived from R-R intervals. To facilitate comparison with other subjects' data, R-R intervals are normalized by dividing each interval by the average R-R interval of that participant. In our experimental study, involving a sample size of N=23, we collected R-R intervals using a Polar H10 heart rate sensor and obtained self-reported pleasantness ratings from cyclists on a scale ranging from 1 (unpleasant) to 3 (pleasant), utilizing an experimental e-bike setup. Our aim is to assess the suitability of heart rate variability analysis for measuring pleasantness. The collected data undergoes preprocessing using two approaches, including a sliding window method. We analyze the data utilizing Support Vector Machines (SVMs) with various kernels and Random Forest classification. The highest achieved F1-score is 47% when using a window size of 130 R-R intervals and incorporating time-domain HRV features such as SDNN, RMSSD, NN50, and pNN50. These findings suggest that additional markers, including frequency-domain HRV and qualitative data, are crucial for a comprehensive assessment. Furthermore, the inclusion of wearable sensor data, such as electrodermal activity, respiration, and skin temperature, may be necessary to accurately estimate self-reported pleasantness in bicyclists.

Additional Key Words and Phrases: heart-rate-variability, bicycling, pleasantness, self-reports, time-domain

1 INTRODUCTION

Heart rate variability analysis plays a crucial role in understanding various parts of the autonomic nervous system [4], which consists of two main divisions: the sympathetic nervous system, and the parasympathetic nervous system. Those divisions work together to regulate involuntary bodily functions like heart rate, blood pressure, respiration, digestion, and sweating. In general, an increase in heart rate variability is seen for positive emotions, whereas decreased heart rate variability is seen for negative emotions.

1.1 Goals, relevance and knowledge gaps

The primary objective of this paper is to assess the applicability of heart rate variability analysis in measuring pleasantness for cyclists. Through our research on this topic, we aim to contribute to a deeper comprehension of the relationship between emotions and their measurable physiological responses. Besides this, if we can obtain quantifiable data that indicates pleasantness for cyclists, methods can be employed to improve the overall cycling experience. The identification of hot spots on a map with negative pleasantness ratings enables road planning authorities to make infrastructure adjustments accordingly. Furthermore, as cycling is an environmentally sustainable mode of transportation, analyzing and improving the overall cycling experience can play a significant role in promoting its adoption. As we will encounter, a lot of research has been done for estimating pleasantness in static conditions through some physiological markers. However, a knowledge gap occurs when exclusively heart rate variability analysis is done in non-static conditions like cycling. The reason for this is that the autonomic nervous system works differently while in exercise: The cardiac parasympathetic- and sympathetic neural activity reduces, which rapidly increases heart rate. During this phase, the sympathetic nervous system works as a tone-setter, and the parasympathetic nervous system works as a rapid modulator [20]. The parasympathetic nervous system plays a role in both pleasant and unpleasant emotions [18]. An even bigger knowledge gap emerges when we want to perform this non-static heart rate variability analysis on an ultra-short-term, so intervals of less than 5 minutes. The reason for wanting to perform heart rate variability analysis on an ultra-short-term is that the pleasantness of the trip often changes in cyclists.[1] In the end, does the exclusive use of heart rate variability features for measuring pleasantness yield significant scores for classification models? If not, can we indicate some reasons behind the lack of accuracy?

1.2 Structure of paper

To achieve these goals, an extensive review of existing literature is performed. To explain the findings accordingly, relevant topics are introduced to the reader. The different domains that classify various metrics of heart rate variability are covered first. Secondly, the 'Circumplex Model of Affect' by Russell [24] is explained to model the subjective concept of pleasantness. Then an experimental study design is composed, explained, and executed, introducing a framework that is focused on measuring stress response during active travel. During the execution of the experimental study, many limitations were established, which we will write an extensive review about. Based on these findings, future work is proposed, meant for any party who is interested in continuing in this field.

2 KEY CONCEPTS AND DEFINITIONS

2.1 Heart rate variability

Heart Rate Variability (HRV) is a measure of the variation in time intervals between consecutive heartbeats or R-peaks, also called

[1]We conclude this from the results obtained by the experimental design. On average, participants cycled for ~53 minutes and gave 30 self-reported pleasantness ratings. On average, 6 were rated as 'unpleasant', 11 as 'neutral' and 13 as 'pleasant'.

The Twente Student Conference on IT, July 7, 2023, Enschede, Overijssel

an R-R interval (Figure 1). However, HRV is not always as simple as the variance of these R-R intervals, since distinct HRV metrics demonstrate different behaviors of the human body. Therefore, HRV can be separated into three main domains: time-domain, frequency-domain, and the non-linear domain. These domains exist to provide specific insights into numerous physiological processes and conditions. An electrocardiogram (ECG) recording is required for determining HRV metrics, and the time-domain and frequency-domain will be explained in a next section. To give some indication of how to interpret an ECG: The small wave before the R peak (P-wave) is the atrial depolarization or activation, so when the heart contracts [8]. Using P waves for time-domain HRV is physiologically the best option since it represents autonomic regulatory effects of heart rate better than using R-peaks [4]. However, in practice, using P waves is technically difficult without using intracardiac atrial electrograms, so R-R intervals are used.

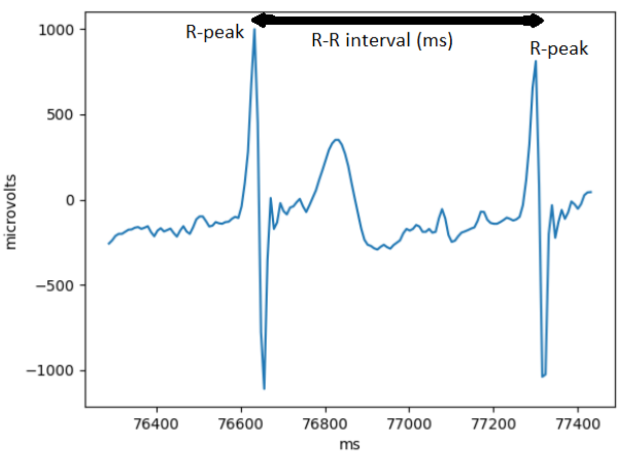

Fig. 1. Electrocardiogram signal, showing the R-peaks and an R-R interval. Source: Collected data from the experimental study.

2.2 Circumplex model of affect

The theoretical basis for using pleasantness as a variable when correlating it to heart rate variability is the 'Circumplex Model of Affect' [24] (Figure 2). This model proposes that all affective states arise from two fundamental neurophysiological systems, one related to valence (displeasure-pleasure) and the other to arousal or alertness (deactivation-activation). Any emotion described by the circumplex model of affect is a linear combination of these two systems [22]. When Valenza et al. (2014) used only ECG data to measure these different neurophysiological systems, they retrieved a higher accuracy on the estimation of the axis of arousal than on the axis of valence [36]. Hence, we are mainly focused on the axis of valence. To measure the axis of valence, it is important to get consistent subjective data from participants when conducting research, which can be done by means of self-reports, which Kalra et al. (2023) wrote a scoping review about [14], or with physiological data, which Lim et al. (2022) wrote a scoping review about [19]. Self-reported pleasantness is difficult to grasp since studies show that people have difficulty in assessing, discerning, and describing their own emotions [22, 25]. Subjects rarely describe that they only feel one emotion, but rather a spectrum of emotions. Preferably, self-reported data needs to come as close to physiological data as possible, but for that, the physiological data needs to be accurate. Lim et al. (2022) questions the validity of self-reports in basic and scientific research since self-reports can be influenced by how interactions are formatted and contextualized [19]. Since the participant is always on the move, there are only a few non-intrusive, less-biased ways of retrieving self-reports. These temporal limitations make it difficult to see how a participant feels throughout the entire ride. The surrounding environment during bicycling and the methods of obtaining self-reports are significant factors to consider in research. These factors also influenced the design of the experimental study conducted in this research.

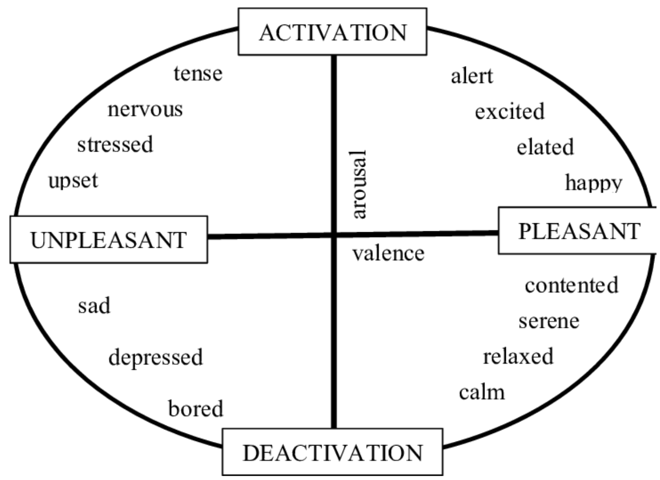

Fig. 2. The Circumplex Model of Affect as defined by J. Posner, J. A. Russell and B. S. Peterson. Source: https://pubmed.ncbi.nlm.nih.gov/16262989/

2.3 Domains of heart rate variability

The different domains of HRV are accurately explained by Shaffer and Ginsberg (2017) and Bilchick and Berger (2006) [4, 32]. We go over the various domains of HRV to comprehend how certain metrics cover different parts of the autonomic nervous system of an individual. To make these domains relevant to our research of measuring momentary pleasantness, we will also touch the concept of short-term HRV metrics, which are HRV metrics calculated on a 5 minute interval.

2.3.1 Frequency-domain. To understand the time-domain metrics, the frequency-domain is covered first. For frequency-domain HRV, most parts of an ECG recording are used, since each part says something different about the influence of the autonomic nervous system. A Fast Fourier Transform (FFT) is applied on an ECG to get the power spectrum, of which frequency bands can be calculated (see Figure 3) [27]. The ultra-low-frequency (ULF, ≤ 0.003 Hz) and very-low-frequency (VLF, $0.003 - 0.04$ Hz) bands are not proficient enough to be used as short-term HRV, since very slow-acting biological and physiological processes like your body clock and metabolism are implicated [2]. The low-frequency (LF, 0.04 − 0.15 Hz) and high-frequency (HF, 0.15−0.40 Hz) bands are more suitable for short-term HRV. LF power is produced by the parasympathetic nervous system (PSNS) and the sympathetic nervous system (SNS), while the HF band power purely reflects PSNS activity. When measuring pleasure, both systems come into play, since the SNS is handling your fight-or-flight response and the PSNS slows down our heart and breathing rate.

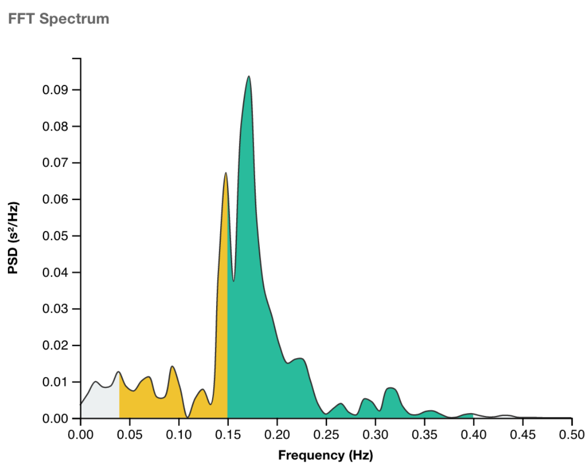

Fig. 3. Power spectral density calculated by a Fast Fourier Transform (FFT). Source: elitehrv.com

2.3.2 Time-domain. Within time-domain HRV, the most important metrics are SDNN (standard deviation of normalized R-R intervals), RMSSD (root mean square of successive differences of normalized R-R intervals), and pNN50 (the percentage of normalized R-R intervals that have a bigger difference than 50ms to the one before it).[2] SDNN exhibits a strong correlation with the ULF, VLF, and HF bands, as both the parasympathetic nervous system (PSNS) and sympathetic nervous system (SNS) contribute to it. On the other hand, RMSSD, which represents the beat-to-beat variance in heart rate, displays a strong correlation with power in the HF band.

2.4 Ultra-short-term heart rate variability

When conducting measurements of ultra-short-term (UST) HRV, it is important to note that even a single erroneous heartbeat can have a significant impact on HRV metrics [33]. Fluctuations in heart rate and R-R intervals are closely linked, but their relationship is not linear. The reciprocal operation does not follow a linear pattern [17]. The reason for discussing UST HRV measurements separately is that while long-term (>60 minutes), short-term (5 minutes), and ultra-short-term (<5 minutes) recordings may employ the same mathematical formulas to calculate HRV metrics, they are not interchangeable.

[2]Normalized R-R intervals are also called N-N intervals, in which outliers are removed. Outliers exist because of inaccuracies in recording equipment. For simplicity, only R-R intervals are mentioned in this paper.

2.4.1 Static conditions. The UST HRV analysis that Shaffer et al. (2016) performed on patients that were in a resting state generally yields a higher accuracy for time-domain HRV than for frequency-domain HRV (See Figure 4) [34]. In Shaffer's tests, the most used HRV metrics and their UST value were compared to the actual HRV of 5-minute mean values. The standard deviation of R-R intervals (SDNN) and the root mean square of subsequent differences (RMSSD) have significant probabilities (~ 80%) of being the true HRV measured on an interval of only 20 seconds. When looking at this probability at a 1-minute recording time, the probability goes up to ~95% for both metrics.

Table 1. Pearson product-moment correlations between UST and 5-minute mean values.

HRV Index	10 s	20 s	30 s	60 s	90 s	120 s	180 s	240 s
Heart rate	[illegible]	.973**	.977**	.986**	.990**	.991**	.996**	.999*
SDNN	.685**	.855**	.897**	.950**	.965**	.963**	.986**	.994**
RMSSD	.597**	.796**	.893**	.945**	.948**	.969**	.988**	.995**
NN50	[illegible]	.653**	.795**	.924**	.948**	.969**	.985**	.992**
pNN50	.653**	[illegible]	.820**	.930**	.947**	.968**	.985**	.994**
HRV triangular index	.389*	.675**	.647	[illegible]	[illegible]	.904	.929**	.925**
TINN	.713**	.795**	.808**	.888**	.906**	.820**	.953**	.972**
VLF power	.320*	.016	.149	.315*	.539**	.526**	.901**	.968**
LF power	.675**	.607**	.795**	.825**	.919**	.934**	.985**	.999**
LFnu	.262	.462**	.550**	.563**	.693**	.774**	.936**	.994**
HF power	.492**	.303	.590**	.746**	.813**	.843**	.974**	.990**
HFnu	.261	.460**	.549**	.563**	.692**	.772**	.936**	.994**
LF/HF power	.026	.108	.349*	.646**	.695**	.800**	.962**	.998**

Fig. 4. Pearson product-moment correlations between UST and 5-minute mean values for participants in a resting state. Source: *Shaffer et al. (2016)* [34]

2.4.2 Non-static conditions. Measuring ultra-short-term HRV for people in the resting state was a reliable tool. However, what about non-static conditions? Kim et al. (2021) conducted an analysis on ultra-short-term HRV while people were in states other than resting, like bicycling [16]. Their experimental protocol included making ECG recordings for 23 minutes that consisted of a resting state, an exercising state, and a post-exercise recovery state. The participant requirements were similar to that of the experimental design of this study, for example, no drugs like alcohol, caffeine, or nicotine were allowed. Based on statistical features, like Cohen's d, Pearson's r, and Limits of Agreements, the researchers determined what HRV metrics are accurate enough to use as features while people are cycling, which can be found in Table 1. Kim et al. (2021) outed two recommendations, one that is more strict and one that is more lenient. In general, a longer recording is required in almost all cases for the ultra-short-term HRV analysis of non-static conditions compared to the static condition. Another interesting conclusion from the paper by Kim et al. (2021) is that the frequency-domain metrics on the HRV spectrum [5] TP (total power), LF (low-frequency band), HF (high-frequency Band), LF/HF (ratio of the LF and HF band) and nHF (normalized high-frequency band) are more qualified as a reliable ultra-short-term HRV metric in non-static conditions.

3 STUDY DESIGN

The focus of this experimental study is to see if the discussed metrics of HRV are relevant for measuring the pleasantness of our data subjects. We will begin by establishing a foundational model. Based

Table 1. Suggested minimum analysis intervals for UST HRV by Kim et al. (2021) [16]

HRV variables (s)		
Metric	Recommendation (strict)	Recommendation (lenient)
SDNN	n/a	180
pNN50	180	10
RMSSD	n/a	10
TP	120	30
VLF	240	120
LF	n/a	30
HF	n/a	30
LF/HF	240	30
nHF	120	30

on this model, specific choices have been made for the study design. We have opted for a within-subject design, where participants experience the same conditions. To differentiate between self-reports and pleasantness, participants were asked thirteen post-ride questions to gain a comprehensive understanding of their overall trip. Although the answers to these questions are not directly used for analysis, they can be used by the researcher to assess the consistency of participants' indications of pleasantness. The questions and answers can be found in Appendix B.1. The eventual goal of using a classification model is to take an arbitrary HRV value, which was determined by a reliable metric for that time span, and use it to guess what a participant would indicate in terms of pleasantness at that time.

3.1 The model

As Bigazzi et al. (2022) created a conceptual framework for measuring stress response during active travel [3], we can use relevant factors, appraisals, responses, and physiological outcomes as a basis for modeling pleasantness. This is because the emotion 'stress' has a linear relation to the axis of valence according to the Circumplex Model of Affect. The framework was created with the idea that research on the topic was plagued by a lack of consistency, and properly shows what is involved around the topic of stressors. The framework consists of four main layers, and each layer is a result of the layer above, eventually leading to physiological outcomes, like HRV.

3.1.1 Factors. Those that influence the overall travel experience in relation to a perception of safety, comfort, physical state, physical demands, or satisfaction. The factors can be split up into two main divisions: environmental factors, that for instance include the weather, the type of road, and other road users. And travel factors, so the equipment weight, helmet usage, and the purpose of the trip.

3.1.2 Appraisal. Each combination of factors (sub-)consciously creates a stimulus for an individual. This stimulus can either be appraised as positive, negative, or irrelevant. Then a balance of personal resources and demands is performed on the stimuli. Are the stimuli familiar or novel? Is it controllable? And is it supportive?

3.1.3 Response. The cognitive-affective and behavioral responses are triggered based on whether the stimulus is perceived as a challenge or a threat. The cognitive-affective response involves determining how to feel in that moment, while the behavioral response involves consciously deciding how to act, such as cycling faster or slower. These responses are crucial for our research as they can result in observable physiological changes related to arousal and/or valence.

3.1.4 Physiological outcomes and markers. Examples of outcomes in physiology are related to the autonomic nervous system and can result in changes in respiration, pupil dilation, heart rate variability, and skin conductance.

3.2 Participant info

In this study, 23 adult participants (aged 18 or older) took part. All participants self-reported good health so with no known heart diseases. They were experienced cyclists, engaging in cycling at least once per week for the past 6 months. Prior to participation, participants abstained from prohibited substances such as alcohol, cannabis, and other non-illicit drugs that could alter heart rate conditions, like caffeine, for 24 hours prior to data collection (Karapetian, 2012) [15]. All participants joined the study at their own risk and were expected to have their own general health insurance. The risk associated with participating in traffic as a cyclist during the data collection period was not different from regular traffic participation. Participants were encouraged to not take any more or less risks than they would on a regular bike ride. An option was provided for participants to be personally informed if there were significant deviations in their heart rate data that were not caused by hardware defects. Each participant received the same instructions, which included reporting their feelings approximately every 3 minutes as indicated by an LED. They were instructed not to cycle significantly faster or slower than their usual pace and to simply enjoy the ride. Additionally, a clear explanation of the term "pleasantness" was provided to ensure consistent interpretation among participants. They were instructed to rate their current pleasantness based on a scale where a low score indicated being upset or sad, the highest score indicated being happy and content, and anything in between (e.g., tense, alert, bored, calm) was considered a neutral score. By providing these instructions and clarifications, the study aimed to maintain consistency in data collection.

3.3 Setting and conditions

The data collection is resolved around field research in a natural setting. Participants were free to choose their own route. Choosing a natural setting enables this variety in environmental factors as shown in the model, enabling diversity for the other layers. Due to the limitation of having a non-water resistant measuring setup, participation was only possible when no precipitation was expected during the trip. The participant was provided an e-bike with full control of motor supports, to make the cycling experience less of an exercise, where heart rate would increase drastically.[3] Incorporating

[3]In the results, we measured a total of 91858 R-R intervals. The average R-R interval was 558ms, which indicates an average heart rate of 108 bpm. The participants of the study by Kim et al. [16] also get similar R-R intervals in the 'exercising' state.

an e-bike in the experimental study would also make people more aware of the factors and appraisals as mentioned in the model since parasympathetic nervous activity would be more dominant than in cases of high-level exercise [16, 20].

3.4 Data collection method

The experimental study was prepared by creating a data collection device mounted on an e-bike. The setup contained 3 buttons and an LED. The 3 buttons represented a scale: unpleasant, neutral, and pleasant, which was indicated by a sad, neutral, and happy emoji. The LED started blinking after 3 minutes of no input from the data subject. Participants were allowed to indicate their pleasure more often, but also less frequently was permitted. Polar Research and Technology conducted an experiment regarding various heart rate sensors that are able to capture ECG images for people at rest, while running, stationary bicycling, and weight training. They captured all these states in the same session. The H10 detects R-R intervals within 2ms, and the accuracy while bicycling is 99.3% [35]. Based on other research by Schaffarczyk et al. (2020), the bias and limits of agreements for measuring R-R and heart rate were minimal, which indicates sufficiency for usage in this research [30]. The data of the Polar H10 was cast by Bluetooth to the 'Polar Sensor Logger' application, which was installed on a separate phone. This phone was either given to the participant or put in the same box as the power bank, Raspberry Pi 4, and cables, to be carried close to the H10. See Figure 5 for the data collection setup.

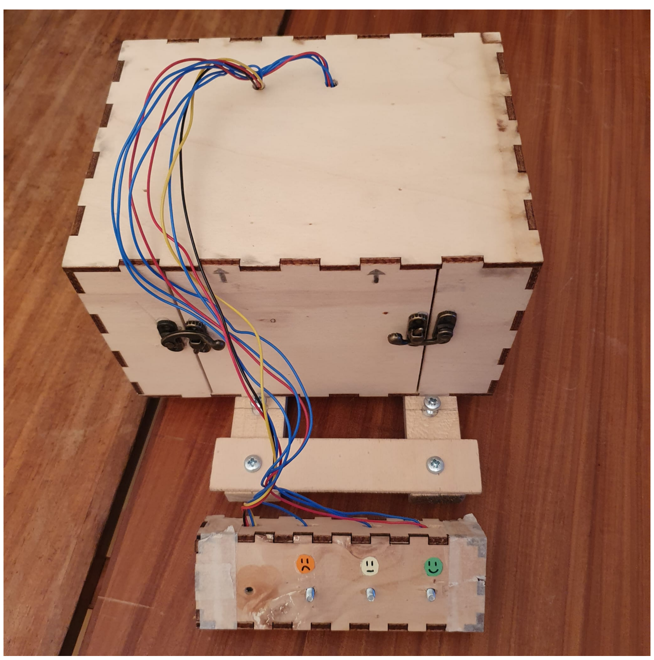

Fig. 5. The data collection setup that was attached to the e-bike. The small user interface was attached to the steering wheel and the frame with equipment was placed at the front of the bike. See Appendix B.2-Figure 9

3.5 Ethical approval

This study design received ethical approval from the Committee of Computer & Information Sciences (CIS). The application number is 230362. Before the data collection, all participants received an information letter by email or by a WhatsApp message. All participants filled in an informed consent form.

3.6 Procedure

The participants could choose their own time slot and the premise of cycling between 6:30 AM and 9:30 AM or 3:30 PM and 7:30 PM was found to be redundant for this experiment. The meet-up place was also not important for the data collection, so the start location was chosen in collaboration with the participant and researcher. The researcher brought the prepared e-bike, with buttons and LED. The researcher made sure that the e-bike was charged so that motor support could be used as the participant desired. The researcher also made sure that the power bank connected to the Raspberry Pi 4 was fully charged. See Figure 6 for an overview of the procedure.

3.7 Pre-processing, classification and feature selection

3.7.1 Collected data. After the data collection, files were renamed from the date and time to an anonymous identifier. Any personally identifiable data is not archived. Data was obtained via two sources:

1. 'Polar Sensor Logger' application for Polar H10 data:
   - (a) Raw R-R intervals in milliseconds with a timestamp.
   - (b) Raw ECG (electrocardiogram) data in micro-volts with a timestamp.
   - (c) Raw ACC (accelerometer) data in micro-g forces (x,y,z) with a timestamp.
   - (d) Raw HR (heart rate) data with a time stamp.
2. Standalone measurement device data
   - (a) Raw self-reports on a scale 1, 2, 3 and a timestamp

Data from the Polar H10 'Polar Sensor Logger' application was recorded with a refresh rate of 100Hz. The R-R interval and HR data files were only updated when the H10 recorded an R-peak. The R-peaks that were used for analysis were the R-R intervals provided by the 'Polar Sensor Logger' application, so they were not manually extracted for the raw ECG data.

3.7.2 Data quality assessment. Due to technical errors, some participant data was lost or deemed unusable in the study. For the first subjects (N=3), there was a deviation in the time synchronization between the heart rate sensor and self-reporting due to a real-time clock issue. The starting time for these participants was manually recorded but insufficient for the classification model. Later, a Python script was used to adjust the time stamps based on the manually recorded data. Two collection moments were conducted by the research supervisor instead of the main researcher. For some participants we experienced issues with the Polar Sensor Logger application (N=3), resulting in the loss of R-R intervals. In two cases, the logger was re-executed, generating separate R-R interval files that were later merged. Data was excluded from analysis for one participant that only had six seconds of data due to an unnoticed logger crashing. There were also two instances where button-related problems occurred, either due to an unplugged jumper cable or

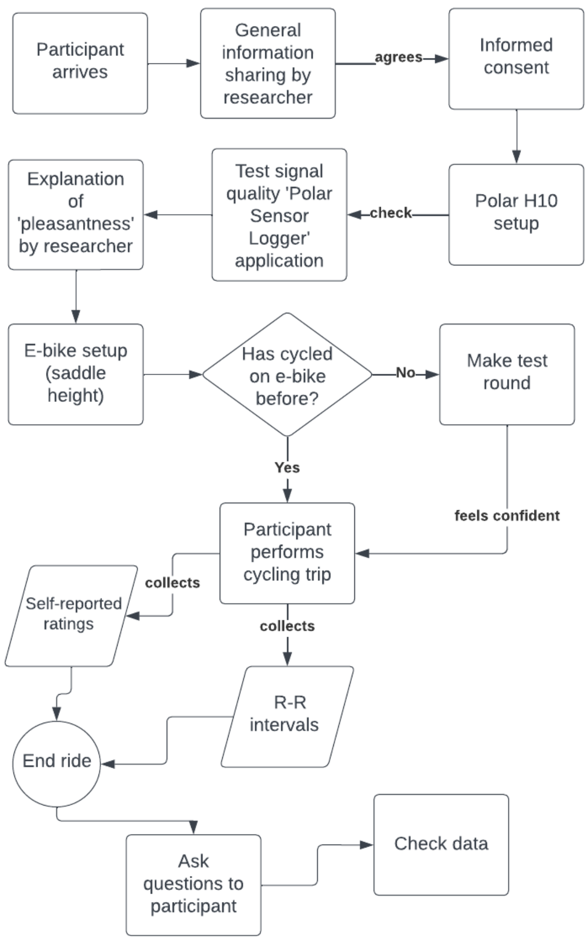

Fig. 6. Flowchart displaying the overall experimental procedure per participant

a misplaced screw, leading to exclusion from the analysis. Out of N=23, only N=17 data entries were used in the analysis.[4] In total, 508 self-reports were extracted, which comes down to an average of 29.88 self-reports per participant. The minimum time of R-R intervals recorded was 34 minutes and 45 seconds. The maximum time of R-R intervals recorded was 1 hour, 33 minutes. The average time that a data subject was participating in the experiment was 52 minutes and 48 seconds. Out of the 508 self-reports given, 98 self-reports were rated as unpleasant, 192 self-reports as neutral, and 218 self-reports as pleasant.

3.7.3 Normalizing R-R intervals. According to Sacha and Pluta (2005, 2008, 2013) [26–28], when using HRV as a dependent continuous variable in research studies, it is vital to consider the non-linear relationship between heart rate (HR) and R-R intervals. Time-domain heart rate variability metrics are influenced not only by autonomic nervous system activity but also by a mathematical phenomenon. Higher HRV is generally associated with lower HR, and vice versa. Therefore, when a participant's heart rate is higher while cycling on an e-bike, their HRV score will automatically be lower. To address this issue of *cycle length dependence* and normalize HRV for all participants, the average heart rate needs to be taken into account. Normalization should be done by dividing each R-R interval by the average R-R interval of that data subject. For a visualization of what happens to the R-R intervals after normalization, see Appendix A.1

3.7.4 Classification model. According to a comprehensive review of wearable-based affect recognition studies conducted by Schmidt et al (2019), support-vector-machine (SVM) classification models are predominantly employed [31]. Interestingly, despite the recognition of random forest classification as the "best family of classifiers" based on the study by Fernández-Delgado et al. [9], ensemble methods like random forest are utilized less frequently for affect recognition. In addition, SVMs offer various kernels, including non-linear ones, which are particularly helpful for our ultra-short-term HRV metrics [6, 10, 32]. Notably, Valenza et al. (2014) demonstrated that achieving up to 95% accuracy in valence classification requires the use of an SVM with a non-linear kernel [36]. SVMs have also been used for stress recognition in numerous studies, i.e. [12, 23]. Consequently, for our classification models, we will utilize both non-linear and linear kernels for SVMs, as well as a random forest classifier.

3.7.5 Feature selection. The study primarily emphasized the normalized R-R intervals obtained from the Polar H10 rather than the individual electrocardiogram (ECG) results. The reliability of the ECG data may be questionable. Specifically, when using the Polar H10 for precise ECG measurements during low and high-intensity activities, experiments by Shaffarczyk et al. (2022) showed that wider bias and limits of agreement were observed [30]. Further details regarding these limitations will be discussed later. Hence, only time-domain features were used for classifying HRV to the target of self-reports. When participants indicated pleasantness through a button press, the corresponding time stamp was recorded, and intervals of varying lengths were calculated around that self-reported time stamp. The interval lengths used for analysis were 40 seconds, 60 seconds, 90 seconds, 120 seconds, and 150 seconds. Various combinations of these interval lengths and four time-domain metrics of HRV were used as features:

(1) SDNN - The standard deviation of R-R intervals.
(2) RMSSD - Root mean square of successive differences of R-R intervals.
(3) NN50 - Number of R-R intervals differing by 50ms or more from its successor.
(4) pNN50 - Ratio of NN50 to the total number of R-R intervals.

Also included was the general HRV over the complete interval of recorded R-R intervals, the reasoning behind this was that if the overall HRV of a person is indicating less or more pleasure, this would also be seen back in the overall trend of the self-reports. Besides this interval selection procedure, an implementation of a sliding window approach was used too, with the same features, but the length of the interval is the window size.

[4] The data points that were excluded from the analysis are #1, #2, #3, #9, #12, and #23.

4 RESULTS

The results are divided into two sections. In the first section, an approach is employed wherein intervals of R-R intervals are established around each recorded self-reported pleasantness time. On this interval, various features of time-domain HRV are calculated. In the second section, a sliding window approach is used, where all R-R intervals are separated into consecutive windows with overlap, and HRV features are calculated on this window. One benefit of utilizing the first approach is the certainty that the data subject experienced the reported pleasantness during that specific moment. However, there are only limited self-reports, which means that the sliding window approach does include all recorded R-R intervals. For all results, training and test sizes were varying from 10% to 50%. Each model was executed 50 times, and an average was taken of the resulting scores.

4.1 Interval around distinct self-reports

Due to the significant class imbalance in the self-reported data used as the target for the classification models, accuracy scores are not suitable and can be unreliable. We have two main ways to counter this. We will publish two results, in one result, the classes will be balanced by excluding 120 pleasant and 94 neutral scores, resulting in 3 classes with 98 self-reports. In the other result, all self-reported times are used and F1-scores will be utilized to demonstrate the results.

4.1.1 Balancing classes. Accuracy scores for the SVM linear kernel, when the classes were balanced, were on average 32%.[5] However, the polynomial and radial basis function SVM kernels reached accuracies of 42.7% and 41.6% respectively. For interval lengths of 30 and 40 seconds, the random forest classification reached 41% accuracy.

4.1.2 F1-scores. When all self-reports were included in the classification model, the linear, polynomial, and RBF kernels did exceptionally worse (<30%). However, the random forest classification reached an f1-score of 46.3% for an interval length of 30 seconds around the self-reported pleasantness time.

4.2 Sliding window

Before we start with stating the results, two main choices had to be made: What window size and overlap should we use and how do we classify each window to a pleasantness rating? Well, each R-R interval is around 550ms on average, which means that a window size of 100 R-R intervals is approximately 55 seconds. We tested various window sizes and their scores for the classification models. For the exact F1 scores per window size for the various classification models, see Appendix C-Figure 11. Each window was classified as a certain pleasantness by looking at the next self-report occurring, any window before this self-report is considered to be rated as the self-reported pleasantness score.

Using the sliding window resulted in less over-/ and underfitting of the models. The SVM linear kernel resulted in bad F1 scores (<32%), while the SVM RBF kernel reached F1 scores higher than chance (36%). The random forest classification worked best and reached a score of 47% for a window size of 130 R-R intervals. Overall, the random forest classification worked best for using HRV features to classify pleasantness. As seen in these results, and as expected by Casties et al. (2006) [6], HRV time-domain metrics cannot be linearly estimated.

[5]Since there are three classifications, this accuracy score is lower than chance.

5 DISCUSSION, LIMITATIONS AND FUTURE WORK

This section will go over the limitations discovered throughout the research. Many points by many different researchers will be discussed, and based on this, a recommendation is made for the interested parties.

5.1 Complexity of HRV

We have discussed that the autonomic nervous system is directly impacting an individual's HRV. However, HRV is also reflected by other systems like the cardiovascular, central nervous, endocrine, and respiratory systems. For instance, measured short-term values of HRV are only appropriate if clients breathe at normal rates (11-20 bpm) since the respiratory system is both influencing and influenced by markers of HRV [32]. Furthermore, as seen by the norms of ultra-short-term HRV measurements, observed values of HRV vary hugely between paced breathing and normal breathing [21]. Subject variables like age, sex, heart rate, and health also have a massive impact on HRV [32]. For future research, including these simple metrics may make classifying pleasantness comprehensible.

5.2 Features

To obtain our results, we specifically concentrated on time-domain HRV features. However, for an activity like cycling, the frequency-domain features are particularly recommended (Table 1). The selected features did not provide sufficient accuracy in predicting the axis of valence in the circumplex model of affect, or in simpler terms, pleasantness. Initially, we excluded frequency-domain features from the experimental design due to the challenges associated with computing the Fast Fourier Transform. We presumed that utilizing ultra-short-term time-domain HRV metrics alone would be adequate to accomplish our objective [16, 33, 34]. Later was established that the Polar H10 heart rate sensor was able to make an ECG for general purposes like acquiring R-R intervals and seeing when the heart contracts. However, using a Polar H10 for getting an ECG that leads to an accurate power spectral density is questionable [30]. To give another example, accurate readings of frequency-domain and non-linear HRV during heavy exercise with a cycloergometer were done by capturing an ECG without a wearable like a Polar H10, but with more accurate equipment [6]. More examples of experiments that measured frequency-domain HRV with more accurate ECG recording equipment are from Barak et al. (2010), Kim et al. (2021), Sacha and Pluta (2005), Salahuddin et al. (2007) and Francesco et al. (2012) [1, 10, 16, 27, 29]. Many experiments that use frequency- or time-domain HRV as an indicator for the modulation of the sympathetic and parasympathetic nervous system or for measuring adaptation of the body in general, use ECG recording lengths of much longer than 300 seconds [1, 5, 6, 38].

According to Bigazzi et al. (2022), the most used physiological marker of stress is electrodermal activity (EDA) [3]. When looking at research that use HRV to measure emotions in the circumplex model of affect (Fujimura and Okanoya, 2012; Valenza et al, 2014; Hovsepian et al. 2015) [11, 13, 36], features other than time-domain HRV are used too. Although we can only speculate about this, extending the feature selection beyond HRV metrics will probably yield better results with respect to measuring the axis of valence. However, since the measurements are done dynamically for bicyclists, the number of wearable sensors used should still be limited because of statistical and logistical reasons [7].

5.3 Classification model

Even though Casties et al. (2006) did not take Sacha's R-R normalization [26] into account for measuring HRV metrics in both resting and exercising state [6], they have concluded that the reliability of using HRV metrics in non-linear classification models still hold. More importantly, our results show that the time-domain HRV features used cannot be linearly estimated, since the radial basis function kernel for the support vector machine classification and random forest classification give higher scores than the linear support vector machine kernel. Furthermore, as shown by Valenza et al. (2012), other non-linear classification models and more features should be used to get higher scores [37]. Considering the findings presented in the results section of this paper, it is advisable to explore person-specific classification models [12]. Numerous studies investigating physiological markers and self-reports in affect recognition often employ classification models tailored to individual data points. This approach is justified given the highly personalized nature of physiological markers [19]. By using person-specific classification models, it becomes possible to account for individual variations and optimize the accuracy of affect recognition. The inclusion of personalized features and characteristics can enhance the precision and reliability of the classification process, leading to more accurate assessments of affective states. However, what remains a challenge is properly classifying the windows based on self-reports, which goes hand in hand with the experimental study setup and simplifying the subjective experience for participants.

5.4 Experimental study

5.4.1 Moment of pleasantness. In the experimental study, an explanation of 'moment' was missing when telling participants to judge their pleasantness at that moment. Participants could have interpreted it as an average pleasantness of the previous 3 minutes, but also as their pleasantness at that exact time, without taking the events that happened in the last 3 minutes into account. Furthermore, no literature could be found about how long pleasantness lasts and how long ECG recordings should be in order to measure HRV accordingly, not even in static conditions. Because of these limitations, window sizes and time interval lengths had no fundamental basis and were 'randomly' approximated.

5.4.2 Participant's bias. Kalra et al. (2023) pointed out many ways to obtain subjective user experiences while riding bicycles [14]. The methods on-ride covered include a push button, sound meter, five-button risk device (similar to ours), and a microphone recording. Also covered are methods for capturing experiences immediately post-ride, but these methods do include a recall bias. However, the 3-minute interval that was incorporated in our study also involves a form of a recall bias, since participants still have to ask themselves how pleasant they feel 'at that moment' based on the events that occurred. A continuous form of capturing subjective experiences would be a microphone recording of the entire ride, in which participants can indicate certain events happening leading to less or more pleasantness. The researcher needs to perform additional pre-processing steps for this type of data capture.

5.4.3 Setting and conditions. The experimental study was designed to be in a natural setting, so without a route chosen by the researcher. Concerns have been raised about the validity of non-natural field setting experiments since it is less accurate in understanding the complex interaction with road infrastructure, motor vehicles, and people's own characteristics. Kalra et al. (2023) state that understanding peoples' subjective experiences while they ride is vital in order to better characterize the factors that contribute to negative riding experiences and identify opportunities to increase bike riding participation [14]. It could be interesting for traffic engineers to see accurate 'pleasantness' data on roads that were constructed, to see if similar design choices should be used. Another approach is to keep the field research in a natural setting, but add GPS data, so that for a high number of data subjects, hot spots can be identified.

5.4.4 Invasiveness of wearables. The measurement setup for the experimental study was created specifically for the purposes of this research. The setup is completely adaptable and may be changed later for further research. Checking the feasibility of creating a standalone setup (Raspberry Pi 4) with hand contact sensors that are able to measure R-R intervals was one of the initial research goals. This goal was dropped because of complexities with hardware and time constraints but is part of the future work since using hand contact sensors is a less invasive way of measuring HRV.

6 CONCLUSION

In this study, valuable insights into measuring the axis of valence in the circumplex model of affect through heart-rate variability have been gained. Using only time-domain HRV metrics in the conditions and setting as we have done for our experimental research is lacking critical markers that estimate pleasantness. When related to HRV features, frequency and non-linear HRV metrics must be included. For even potentially higher accuracy, as seen in other research experiments that estimate pleasantness, non-HRV metrics should be included. These could be physiological markers (EDA, respiration rate, body temperature) but could also be values of other factors that influence pleasantness. In future work, the main focus should be implementing feature-domain HRV features and other non-HRV physiological markers.

REFERENCES

[1] Otto F. Barak, Djordje G. Jakovljevic, Jelena Z. Popadic Gacesa, Zoran B. Ovcin, David A. Brodie, and Nikola G. Grujic. 2010. Heart Rate Variability Before and After Cycle Exercise in Relation to Different Body Positions. *Journal of Sports Science & Medicine* 9, 2 (June 2010), 176–182. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3761735/

[2] Brayans Becerra-Luna, Raúl Cartas-Rosado, Juan Carlos Sánchez-García, Raúl Martínez-Memije, Oscar Infante-Vázquez, Claudia Lerma, Héctor Pérez-Grovas, José Manuel Rodríguez-Chagolla, Gabriel Sánchez-Pérez, and Miguel Cadena-Méndez. 2020. Heart Rate Variability Assessment Using Time–Frequency Analysis

in Hypotensive and Non-Hypotensive Patients in Hemodialysis. *Applied Sciences* 10, 17 (Jan. 2020), 6074. https://doi.org/10.3390/app10176074 Number: 17 Publisher: Multidisciplinary Digital Publishing Institute.
[3] Alexander Bigazzi, Fajar Ausri, Luke Peddie, Dillon Fitch, and Eli Puterman. 2022. Physiological markers of traffic-related stress during active travel. *Transportation Research Part F: Traffic Psychology and Behaviour* 84 (Jan. 2022), 223–238. https://doi.org/10.1016/j.trf.2021.12.003
[4] Kenneth C. Bilchick and Ronald D. Berger. 2006. Heart rate variability. *Journal of Cardiovascular Electrophysiology* 17, 6 (June 2006), 691–694. https://doi.org/10.1111/j.1540-8167.2006.00501.x
[5] Robert L. Burr. 2007. Interpretation of Normalized Spectral Heart Rate Variability Indices In Sleep Research: A Critical Review. *Sleep* 30, 7 (July 2007), 913–919. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1978375/
[6] J.-F. Casties, D. Mottet, and D. Le Gallais. 2006. Non-linear analyses of heart rate variability during heavy exercise and recovery in cyclists. *International Journal of Sports Medicine* 27, 10 (Oct. 2006), 780–785. https://doi.org/10.1055/s-2005-872968
[7] Jongyoon Choi, Beena Ahmed, and Ricardo Gutierrez-Osuna. 2012. Development and evaluation of an ambulatory stress monitor based on wearable sensors. *IEEE transactions on information technology in biomedicine: a publication of the IEEE Engineering in Medicine and Biology Society* 16, 2 (March 2012), 279–286. https://doi.org/10.1109/TITB.2011.2169804
[8] Steven Douedi and Hani Douedi. 2023. P wave. In *StatPearls*. StatPearls Publishing, Treasure Island (FL). http://www.ncbi.nlm.nih.gov/books/NBK551635/
[9] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research* 15, 1 (Jan. 2014), 3133–3181.
[10] Buccelletti Francesco, Bocci Maria Grazia, Gilardi Emanuele, Fiore Valentina, Calcinaro Sara, Fragnoli Chiara, Maviglia Riccardo, and Franceschi Francesco. 2012. Linear and Nonlinear Heart Rate Variability Indexes in Clinical Practice. *Computational and Mathematical Methods in Medicine* 2012 (2012), 219080. https://doi.org/10.1155/2012/219080
[11] Tomomi Fujimura and Kazuo Okanoya. 2012. Heart Rate Variability Predicts Emotional Flexibility in Response to Positive Stimuli. *Psychology* 03, 08 (Aug. 2012), 578. https://doi.org/10.4236/psych.2012.38086 Number: 08 Publisher: Scientific Research Publishing.
[12] Javier Hernandez, Rob R. Morris, and Rosalind W. Picard. 2011. Call Center Stress Recognition with Person-Specific Models. In *Affective Computing and Intelligent Interaction (Lecture Notes in Computer Science)*, Sidney D'Mello, Arthur Graesser, Björn Schuller, and Jean-Claude Martin (Eds.). Springer, Berlin, Heidelberg, 125–134. https://doi.org/10.1007/978-3-642-24600-5_16
[13] Karen Hovsepian, Mustafa al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment. *Proceedings of the ... ACM International Conference on Ubiquitous Computing . UbiComp (Conference)* 2015 (Sept. 2015), 493–504. https://doi.org/10.1145/2750858.2807526
[14] Anjani Kalra, Tommy Lim, Lauren Pearson, and Ben Beck. 2023. Methods used to capture subjective user experiences in adults while riding bicycles: a scoping review. *Transport Reviews* 43, 3 (May 2023), 453–477. https://doi.org/10.1080/01441647.2022.2123064 Publisher: Routledge _eprint: https://doi.org/10.1080/01441647.2022.2123064.
[15] G. K. Karapetian, H. J. Engels, K. A. Gretebeck, and R. J. Gretebeck. 2012. Effect of caffeine on LT, VT and HRVT. *International Journal of Sports Medicine* 33, 7 (July 2012), 507–513. https://doi.org/10.1055/s-0032-1301904
[16] Jin Woong Kim, Hyeon Seok Seok, and Hangsik Shin. 2021. Is Ultra-Short-Term Heart Rate Variability Valid in Non-static Conditions? *Frontiers in Physiology* 12 (2021). https://www.frontiersin.org/articles/10.3389/fphys.2021.596060
[17] Robert Kleiger, Phyllis Stein, M Bosner, and Jeff Rottman. 1992. Time-Domain Measurements of Heart Rate Variability. *Cardiology clinics* 10 (Sept. 1992), 487–98.
[18] Sylvia D. Kreibig. 2010. Autonomic nervous system activity in emotion: A review. *Biological Psychology* 84, 3 (July 2010), 394–421. https://doi.org/10.1016/j.biopsycho.2010.03.010
[19] Tommy Lim, Anjani Kalra, Jason Thompson, Joanne Caldwell Odgers, and Ben Beck. 2022. Physiological measures of bicyclists' subjective experiences: A scoping review. *Transportation Research Part F: Traffic Psychology and Behaviour* 90 (Oct. 2022), 365–381. https://doi.org/10.1016/j.trf.2022.09.007
[20] Scott Michael, Kenneth S. Graham, and Glen M. Davis. 2017. Cardiac Autonomic Responses during Exercise and Post-exercise Recovery Using Heart Rate Variability and Systolic Time Intervals—A Review. *Frontiers in Physiology* 8 (2017). https://www.frontiersin.org/articles/10.3389/fphys.2017.00301
[21] David Nunan, Gavin R. H. Sandercock, and David A. Brodie. 2010. A Quantitative Systematic Review of Normal Values for Short-Term Heart Rate Variability in Healthy Adults. *Pacing and Clinical Electrophysiology* 33, 11 (2010), 1407–1417. https://doi.org/10.1111/j.1540-8159.2010.02841.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-8159.2010.02841.x.
[22] Jonathan Posner, James A. Russell, and Bradley S. Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology* 17, 3 (2005), 715–734. https://doi.org/10.1017/S0954579405050340
[23] Pramila Rani, Changchun Liu, Nilanjan Sarkar, and Eric Vanman. 2006. An Empirical Study of Machine Learning Techniques for Affect Recognition in Human–robot Interaction. *Pattern Analysis and Applications* 9 (May 2006). https://doi.org/10.1007/s10044-006-0025-y
[24] James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (Dec. 1980), 1161–1178. https://doi.org/10.1037/h0077714
[25] Carolyn Saarni. 1999. *The development of emotional competence*. Guilford Press, New York, NY, US. Pages: xviii, 381.
[26] Jerzy Sacha. 2013. Why should one normalize heart rate variability with respect to average heart rate. *Frontiers in Physiology* 4 (Oct. 2013), 306. https://doi.org/10.3389/fphys.2013.00306
[27] Jerzy Sacha and Władysław Pluta. 2005. Different methods of heart rate variability analysis reveal different correlations of heart rate variability spectrum with average heart rate. *Journal of Electrocardiology* 38, 1 (Jan. 2005), 47–53. https://doi.org/10.1016/j.jelectrocard.2004.09.015
[28] Jerzy Sacha and Władysław Pluta. 2008. Alterations of an average heart rate change heart rate variability due to mathematical reasons. *International Journal of Cardiology* 128, 3 (Aug. 2008), 444–447. https://doi.org/10.1016/j.ijcard.2007.06.047
[29] Lizawati Salahuddin, Jaegeol Cho, Myeong Gi Jeong, and Desok Kim. 2007. Ultra Short Term Analysis of Heart Rate Variability for Monitoring Mental Stress in Mobile Settings. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 4656–4659. https://doi.org/10.1109/IEMBS.2007.4353378 ISSN: 1558-4615.
[30] Marcelle Schaffarczyk, Bruce Rogers, Rüdiger Reer, and Thomas Gronwald. 2022. Validity of the Polar H10 Sensor for Heart Rate Variability Analysis during Resting State and Incremental Exercise in Recreational Men and Women. *Sensors* 22, 17 (Aug. 2022), 6536. https://doi.org/10.3390/s22176536
[31] Philip Schmidt, Attila Reiss, Robert Dürichen, and Kristof Van Laerhoven. 2019. Wearable-Based Affect Recognition—A Review. *Sensors* 19, 19 (Jan. 2019), 4079. https://doi.org/10.3390/s19194079 Number: 19 Publisher: Multidisciplinary Digital Publishing Institute.
[32] Fred Shaffer and J. P. Ginsberg. 2017. An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in Public Health* 5 (2017). https://www.frontiersin.org/articles/10.3389/fpubh.2017.00258
[33] Fred Shaffer, Zachary M. Meehan, and Christopher L. Zerr. 2020. A Critical Review of Ultra-Short-Term Heart Rate Variability Norms Research. *Frontiers in Neuroscience* 14 (2020). https://www.frontiersin.org/articles/10.3389/fnins.2020.594880
[34] Fredric Shaffer, Steven Shearman, and Zachary Meehan. 2016. The Promise of Ultra-Short-Term (UST) Heart Rate Variability Measurements. *Biofeedback* 44 (Dec. 2016), 229–233. https://doi.org/10.5298/1081-5937-44.3.09
[35] Polar Research and Technology. 2019. Polar H10 Heart Rate Sensor System. https://www.polar.com/en/img/static/whitepapers/pdf/polar-h10-heart-rate-sensor-white-paper.pdf
[36] Gaetano Valenza, Luca Citi, Antonio Lanatá, Enzo Pasquale Scilingo, and Riccardo Barbieri. 2014. Revealing Real-Time Emotional Responses: a Personalized Assessment based on Heartbeat Dynamics. *Scientific Reports* 4, 1 (May 2014), 4998. https://doi.org/10.1038/srep04998 Number: 1 Publisher: Nature Publishing Group.
[37] Gaetano Valenza, Antonio Lanata, and Enzo Pasquale Scilingo. 2012. The Role of Nonlinear Dynamics in Affective Valence and Arousal Recognition. *IEEE Transactions on Affective Computing* 3, 2 (April 2012), 237–249. https://doi.org/10.1109/T-AFFC.2011.30 Conference Name: IEEE Transactions on Affective Computing.
[38] V. Vesterinen, K. Häkkinen, E. Hynynen, J. Mikkola, L. Hokka, and A. Nummela. 2013. Heart rate variability in prediction of individual adaptation to endurance training in recreational endurance runners. *Scandinavian Journal of Medicine & Science in Sports* 23, 2 (March 2013), 171–180. https://doi.org/10.1111/j.1600-0838.2011.01365.x

A R-R INTERVALS

A.1 Normalization of R-R intervals

R-R intervals can also be referred to as N-N intervals, which are the R-R intervals where outliers are removed. An outlier was classified as an arbitrary R-R interval that has a relative difference of at least 150ms to any of its 5 adjacent R-R intervals. For simplicity, only R-R intervals are mentioned throughout the research.

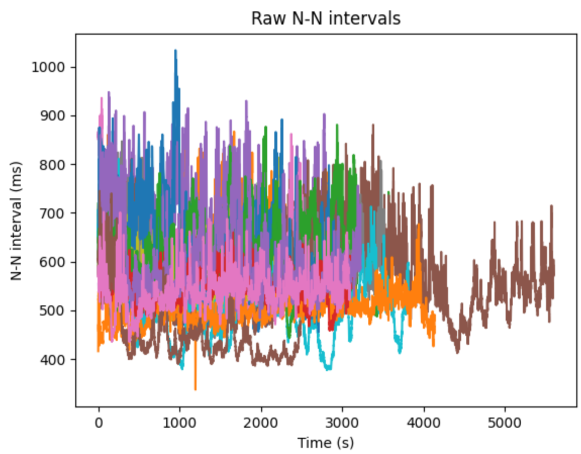

Fig. 7. R-R intervals from the experiment, not yet normalized

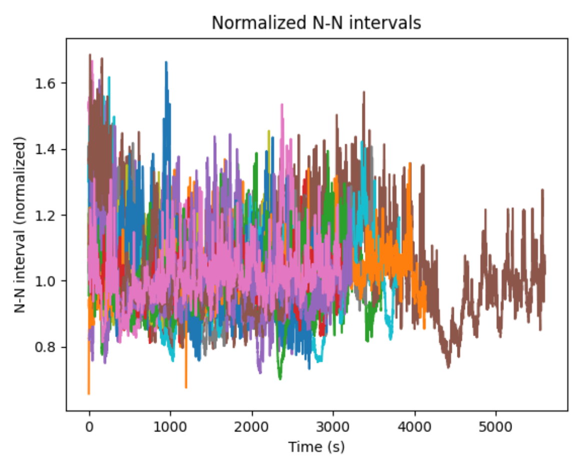

Fig. 8. R-R intervals from the experiment normalized by the way Sacha (2013) described [26].

B EXPERIMENTAL STUDY

B.1 Qualitative questions and answers

The questions asked to each participant after the ride. Participants were asked to rate these statements on a scale from 1 to 5. (Table B.1.1)

(1) Satisfaction during the ride. (Did everything go according to plan?)
(2) Sad- to gladness during the ride. (Were you disappointed or excited?)
(3) Happiness during the ride. (Or were you depressed?)
(4) The ride exceeded all my expectations.
(5) The traffic flow during my journey was smooth and efficient.
(6) The route was enjoyable.
(7) The roads provided a smooth and enjoyable biking experience.
(8) Navigating was effortless.
(9) Overtaking other cyclists was effortless.
(10) Navigating the road became challenging due to multiple difficulties with other road users.
(11) There were numerous obstacles (such as bumps, poles, etc.).
(12) The wind conditions were ideal.
(13) There were too many other road users.

#	1	2	3	4	5	6	7	8	9	10	11	12	13
1	4	3	4	4	5	4	3	5	n/a	1	2	5	3
2	4	3	4	3	5	2	1	5	4	2	3	4	4
3	4	5	4	3	2	4	3	5	4	3	2	4	5
4	5	5	5	5	3	5	1	1	n/a	1	5	5	4
5	3	4	4	4	5	5	4	5	5	2	2	3	4
6	3	4	4	3	2	4	3	5	4	2	2	5	2
7	5	4	5	4	3	3	4	5	5	2	4	5	3
8	4	5	5	3	3	3	4	2	5	4	4	4	3
9*													
10	4	3	5	3	2	3	2	5	5	5	5	5	3
11*													
12*													
13	5	4	4	3	4	3	4	5	4	4	3	2	3
14	4	5	5	3	4	4	4	5	5	2	3	4	2
15	3	4	5	2	4	4	2	3	5	2	4	2	2
16	3	5	5	4	5	4	3	5	4	3	2	1	3
17	4	3	3	3	4	4	5	2	5	2	3	5	2
18	5	4	5	3	3	4	3	5	5	3	3	3	4
19	4	4	3	3	4	4	4	4	5	2	1	3	2
20	3	4	4	2	2	3	2	1	3	3	4	5	3
21	3	3	3	2	4	4	2	5	3	2	3	5	4
22	4	4	4	3	3	3	3	4	3	2	2	4	2
23	3	3	4	3	2	4	3	4	4	3	3	2	4

Table B.1.1: Questions and answers by N=23 participants on 13 questions.

*Rows are empty because either the main researcher was not present, or the participant did not want to participate.

B.2 Experimental setup

Fig. 9. Data collection setup attached to the e-bike.

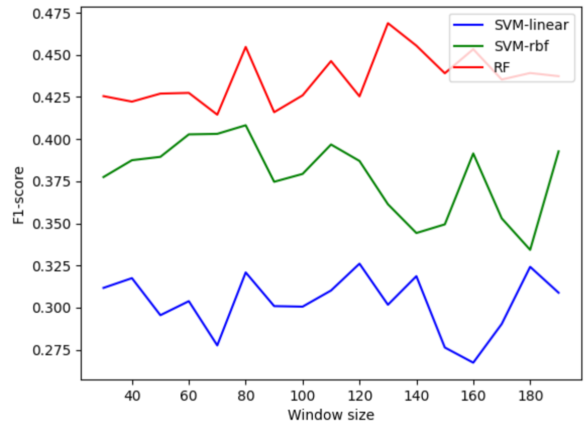

Fig. 11. Various windows and F1-scores for the different classification models

C RESULTS

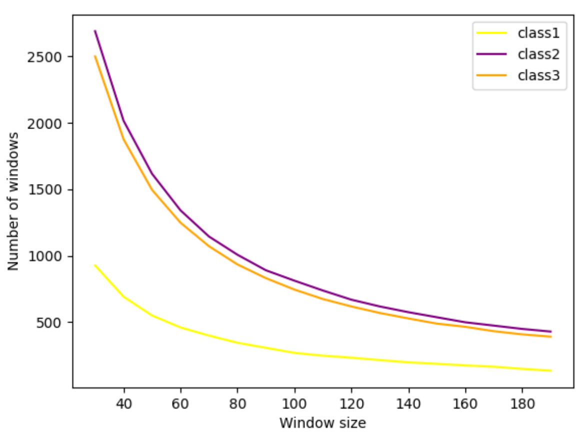

Fig. 10. The number of windows related to the various window sizes of lengths between 30 and 200 R-R intervals.