Cognitive State Classification Using Psychophysiological Measures

Masterarbeit

vorgelegt von

Philipp von Bauer

an der

Universität
Konstanz

Arbeitsgruppe Mensch-Computer-Interaktion

Fachbereich Informatik

1.Gutachter: Prof. Dr. Harald Reiterer
2.Gutachter: Dr.-Ing. habil. Christian Borgelt

Konstanz, 2018

Selbstständigkeitserklärung

Ich versichere, dass ich die beiliegende Master Arbeit selbstständig verfasst und keine anderen Hilfsmittel als die angegebenen benutzt habe. Die Stellen, die anderen Werken dem Wortlaut oder dem Sinn nach entnommen sind, habe ich in jedem einzelnen Fall durch Angaben der Quelle, auch der benutzten Sekundärliteratur, als Entlehnung kenntlich gemacht.

Ort, Datum:

Unterschrift:

Author's Previous Work

Some of the content, figures and tables are based and/or taken from the *Seminar for the Master Project* [73] and from the *Report of the Master Project* [74]:

Master Seminar:
Philipp von Bauer, November 2017
Towards real-time cognitive state assessment using psychophysiological measures

Master Project:
Philipp von Bauer, April 2018
Towards Classifying Cognitive State

Cognitive State Classification Using Psychophysiological Measures

Master Thesis

University of Konstanz
Department of Computer and Information Science
Chair of Human-Computer Interaction, Prof. Dr. Harald Reiterer

Philipp von Bauer
philipp.bauer@uni-konstanz.de
01/950207

First Examiner	Second Examiner
Prof. Dr. Harald Reiterer	Dr.-Ing. habil. Christian Borgelt

Supervisors
Johannes Zagermann & Dr. Ulrike Pfeil

September 12, 2018

Abstract. Assessing users' cognitive state is one of the visions in Human-Computer Interaction (HCI). On the one hand it would allow the building of intelligent adaptive systems and on the other hand, could serve as an in-place evaluation tool of workload in contrast to the current practice of using questionnaires post hoc. Building adaptive human-computer systems is a challenging task. This work addresses one aspect of it by investigating the use of psychophysiological measures - pupil diameter, heart rate and skin conductance - as input for supervised machine learning to classify the cognitive state of users. From 24 subjects data was collected during the n-back task - an abstract working memory task. A Random Forest Classifier (RFC) was trained across users using statistical features of the pupil size. Classification accuracies reached up to 89% in discriminating between the 1-back and 2-back task with a window size of 60 seconds. The other physiological measures were not sensitive to the task's manipulation, thus, left out for classification.

Cross-user classification yielded promising results for the n-back task with letters. In a second explorative small-scale study, six participants performed the n-back task with three new stimuli types instead of letters - spatial, audio and images. The attempt to use the classifier of the first study for cross-task classification showed promising results with accuracies between 79% and 94%.

The cross-user and cross-task classification performed well showing the feasibility of using pupil measures for classification. Accounting for individual differences needs to be addressed further along with a methodology to evaluate classifier performance beyond using task difficulty which cannot be reliably and equally created across tasks.

Some related work investigated the use of EEG for cross-task and cross-user classification. However, most focus was on building individual models. Thus, this work investigated the use of the pupil diameter for cross-task and cross-user classification using the n-back task and variations of it as related work did for EEG.

Table of Contents

Table of Contents
List of Figures
List of Tables
1 Introduction 1
2 Theoretical Foundations 5
2.1 Cognitive State 5
2.1.1 Working Memory 6
2.1.2 Cognitive Load 6
2.1.3 Summary 7
2.2 Measuring the Cognitive State 7
2.2.1 Psychophysiology 7
2.2.2 Psychophysiological Measures 8
2.2.3 Summary 13
2.3 Related Work 13
2.3.1 Dual-Task & Single-Task 14
2.3.2 Adaptation the Goal, Cognitive State Assessment the Challenge 15
2.3.3 Machine Learning - The Assessment Tool 17
2.3.4 Summary 22
2.4 Modulating the Cognitive State 23
2.4.1 Working Memory Tasks 24
2.4.2 The n-back Task 26
2.4.3 Summary 27
2.5 Summary 27
3 Research Question 29
4 Data Gathering Study 31
4.1 Experiment 31
4.1.1 Design & Variables 31
4.1.2 Procedure 31
4.1.3 Task 33
4.1.4 Apparatus & Data Recording 34
4.1.5 Setting 36
4.1.6 Pretest & Pilot Study 37
4.1.7 Summary 37
4.2 Results & Analysis 37
4.2.1 Analysis Procedure 38
4.2.2 Demographics 39
4.2.3 NASA-TLX, Performance 40
4.2.4 Post Interviews 42
4.2.5 First Glimpse at the Physiological Data 44

4.2.6 Pupil Measures 46
4.2.7 Cardiovascular Measures 48
4.2.8 Electrodermal Activity 50
4.2.9 Discussion & Summary 51
4.3 Summary 52
5 Classifying Cognitive State Across Users 53
5.1 Random Forest Classifier 53
5.2 Goals & Approaches 55
5.3 Preprocessing 57
5.4 Learning 59
5.5 Results 60
5.5.1 Goal 1: Specificity 61
5.5.2 Goal 2: Window Size 63
5.5.3 Goal 3: Individual Differences 64
5.5.4 Goal 4: Utilizing the pupil 65
5.5.5 Goal 5: Validation 66
5.5.6 Individual Models 67
5.5.7 Best Performing Pipelines 69
5.6 Summary 71
6 Classifying Cognitive State Across Tasks 73
6.1 Experiment 73
6.1.1 N-back Task Variations 73
6.1.2 Adjusted Design & Procedure 74
6.1.3 Pilot Study 75
6.1.4 Summary 76
6.2 Results & Analysis 76
6.2.1 NASA-TLX, Performance, Post Interviews 76
6.2.2 Pupil Analysis 79
6.2.3 Summary 81
6.3 Classifying Cognitive State: Across Task and Users 81
6.3.1 Preprocessing, Choice of Population Model, Data Scaling 82
6.3.2 Results: *CALIB* and *SPT* 83
6.4 Summary 84
7 Summary & Discussion 85
7.1 Summary & Contribution 85
7.2 Sensitivity of EDA & BVP Measures 86
7.3 2-back, 3-back & the Pupil 87
7.4 Research Question 1: Cross-User Classification 88
7.5 Research Question 2: Cross-Task Classification 90

7.6 Related Work 91
7.7 Limitations 93
8 Conclusion 95
References 98
A Performance, TLX Study 1 104
B Pupil Analysis Study 1 106
C EDA Analysis Study 1 113
D BVP Analysis Study 1 120
E Cognitive State Classification Study 1 126
F Performance Study 2 146
G Contents of the USB flash drive 147

List of Figures

1 Adaptive System 2
2 Psychophysiology 8
3 EDA Signal 11
4 n-back With Letters 26
5 Trial Run 33
6 Data Collection 35
7 Study Setting 36
8 Machine Learning Preprocessing 58
9 Specificty Bar Chart 61
10 Confusion Matrix 3-Class 62
11 Window Size 63
12 Individual Differences 64
13 Features 66
14 N-back Variation Stimuli 74
15 Study 2 Procedure 75
16 Example Pupil Data 79

List of Tables

1 Psychological Constructs 5
2 Working Memory Tasks 25
3 Correlation Matrix Data Gathering Study 40
4 NASA-TLX Descriptives 41
5 NASA-TLX, Performance Post Hoc Tests 41
6 Performance Descriptives 42
7 Pupil Post Hoc Tests 48
8 EDA Descriptives 50
9 Specificity Descriptives 61
10 Individual Models Descriptives 68
11 Pipeline Selection 70
12 Pipeline Selection 2 71
13 NASA-TLX n-back Variants 77
14 N-back Variants Perceived Difficulty 78
15 Pupil Descriptives n-back Variants 80
16 Pupil Variants' Differences 81
17 Classification Accuracies n-back Variants 83

1 INTRODUCTION

One core aspect of Human-Computer Interaction (HCI) is to enhance the experiences users have when interacting with technology. While for many years improving usability was the focus of enhancements, today also hedonic qualities are important - the user experience (UX) [68]. In product development a good UX can be created by using user-centered design (UCD) which Oviatt [57] describes as an approach to

> "model users' natural behaviour [...], including constraints on their ability to attend, learn, and perform, so that interfaces can be designed that are more intuitive, easier to learn, and freer of performance errors." [57]

This description presents the modelling of users as a way to inform the design of interfaces. If it is possible to model the user in real-time during the interaction with a system, new possibilities open up as the design can be adapted to users' needs. There are several benefits of so-called adaptive systems (AS) such as a potentially better user performance [63] which can be seen as supporting the achievement of users' goals [1] along with a reduction of users' workload [25].

Figure 1 shows the three steps of an adaptive system - perceive, select and act. As an example imagine the situation where a driver of a car is on the way home. After a long day, he is exhausted and not able to focus on the road. The system could detect this (perceive) and could decide to take over some functions such as speed control and the brakes (select and act). Thus, the adaptive system reduces the likelihood of an accident.

In order to develop such a system, a reliable detection of the driver's state is required. Hence, the challenges of the perceive-step need to be addressed. One of these is formulated by Feigh et al. [25] as

> "the need for more robust, accurate, wearable, and unobtrusive neurological and physiological sensors capable of providing the real-time information needed to determine users's cognitive state."

This challenge motivates and informs the topic of this thesis: classifying users' cognitive state by using psychophysiological measures. Not only the field of adaptive systems can benefit from the cognitive state classification as it can as well be used as an in-place evaluation tool of workload in contrast to the current practice of using post hoc questionnaires (e.g. NASA-TLX [33]).

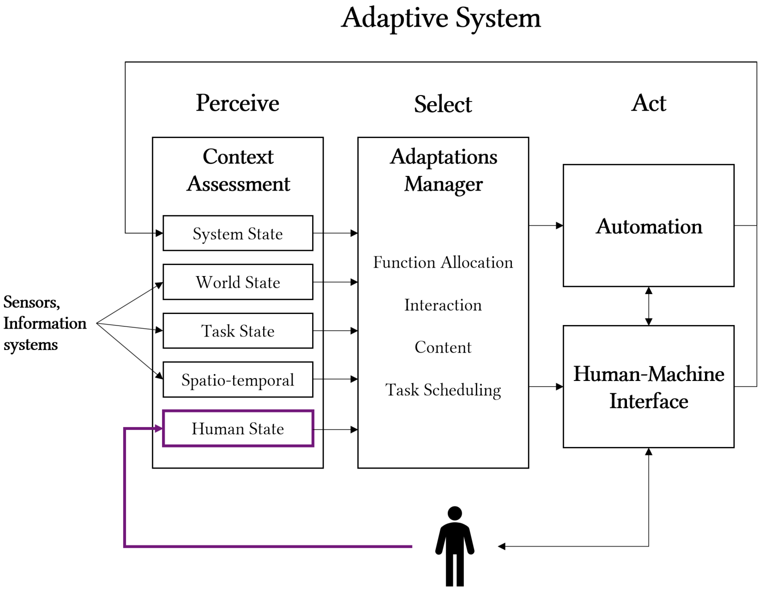


Fig. 1. Generic Adaptive Human-Machine System adopted from [25]

The focus of the thesis is highlighted, namely, perceiving the human state; in particular the cognitive state.

The thesis' content builds up as follows:
Section 2 introduces the fundamental theoretical background information and related work as described in the following.

First, it is required to understand humans' cognitive state and processing. Different theories and constructs from psychology have been used in HCI literature such as cognitive load theory (CLT) [71] or mental effort - these will be subject of section 2.1.

Second, having an understanding of these will raise the question of how these constructs can be manifested in the real world, in other words, measured - section 2.2. The quote by Feigh et al. [25] already mentions a possibility, namely, physiological sensors. The latter allows for an objective method to assess the cognitive state that as well yields the potential to be applied in real-time.

Third, with the basics understood, section 2.3 reviews related work to learn from the attempts already made to assess the cognitive state. Supervised machine learning was frequently used as a tool for the state assessment. Physiological example data to learn from and to predict the state of the user are required. Thus, in order to implement this approach, ground truth data needs to be gathered. Therefore, fourth, the question of how cognitive state can be modulated reliably to obtain this data is subject of section 2.4.

Based on the theoretical background, the research questions this work addresses are presented in section 3. The first question asks if an abstract working memory task can be used to elicit physiological responses that can be fed to a machine learning classifier to predict users cognitive state across individuals. An experiment - section 4 - was conducted gathering physiological data which was then used to build machine learning models - section 5. The second experiment - section 6 - was conducted to explore if the model created with the first experiment's task is also able to predict the user state of variations of the task - section 6.3.
The results of both experiments that addressed the research questions are discussed in section 7 before this work will give a conclusion and outlook to future research.

2 THEORETICAL FOUNDATIONS

This section presents the theoretical background information and related work. It was stated previously that there is a need to have an understanding of the human cognitive state (section 2.1) and how it can be measured using psychophysiological measures (section 2.2). The first two subsections of the theoretical foundations' section are going to address both. By doing so it the cognitive state and the meaning it refers to in the context of the thesis is explained. Further, the physiological measures most relevant, due to their use in the experiment, are introduced. This fundamental knowledge allows for a discussion of related work of the field of HCI where adaptation or cognitive state assessment is subject of interest (section 2.3). Working memory tasks that are commonly used to produce different cognitive states are discussed (section 2.4). The discussion of related work highlights the approaches that have been made already and thus informs the choice of the research question which is presented in section 3.

2.1 Cognitive State

In this section different theories and models from psychology are presented which can be used to model the cognitive state of users. For this work, it is necessary to understand what is considered the cognitive state.

In HCI literature different terms are used to describe the cognitive state. Some examples are cognitive workload [1], mental workload [50], cognitive load [16] or workload including physical effort [34], mental effort [7] or taskload [64]. While all are based on psychology the common ground of these is the working memory with its limited capacity [16]. In table 1 a summary of the terms discussed in this thesis is given.

Construct	Definition	Remarks
Working Memory (WM)	Storage of conscious information and processor of information (organising, contrasting, comparing, working). Limited regarding the amount of information it can hold and process	Miller [48]: 7 (+-2) chunks of information can be hold independent of type Baddeley [5]: Two types of information (phonological loop, visuospatial sketchpad) taking up different capacities. Exceeding capacity is undesired resulting in decreases in performance. Working memory is fundamental for the other constructs.
Cognitive Load (CL)	The load put on the WM during performing a cognitive task. Three components of the CL are mental effort, mental load and performance.	CL Theory categorizes the load into three different types (intrinsic, extraneous, germane).
Mental Load (ML)	The relation of the task demands and the individual's ability to meets these demands.	-
Mental Effort (ME)	The amount of resources allocated for the current task.	It might be seen as the actual measure of used cognitive capacity [71].

Table 1. Overview of the psychological constructs

2.1.1 Working Memory

The working memory (WM) holds the information that is currently being processed [71]. It is part of the cognitive architecture which also includes long-term memory, schema acquisition and automation. Before storing information in the long-term memory (LTM), it is processed in the working memory. Schema theory is concerned with the process of transferring information from the WM to the LTM. A detailed description of the whole architecture and its functioning can be found in [71]. This research focuses on the WM and its aforementioned limited capacity. According to Miller [48] only 7+-2 chunks of information can be kept in the WM. A chunk has no fixed size, and the capacity is also used by the organising, contrasting and processing of information - the main functions of the WM. Miller's [48] theory is a unitary system and for instance, does not take into account the types of information. Baddeley's [5] model consists of two parts: the *phonological loop* and the *visuospatial sketchpad* which are controlled by a *central executive*. The latter as a result controls the processing of information in general. The loop holds all kinds of speech-based information while the sketchpad deals with spatial and visual information.

The working memory is highly important to process information for short-term or long-term use. It can be seen as the *Interface between Memory and Cognition* [5]. Assessing how much of its capacity is in use can be seen as measuring the cognitive state.

2.1.2 Cognitive Load

The term cognitive load (CL) is rooted in the theory by Sweller et al. [71]. Paas et al. [58] went for the following definition:

> "... a multidimensional construct representing the load that performing a particular task imposes on the learner's cognitive system." [58]

Further, the load is described by Sweller et al. [71] as an interaction between three parts: mental load, mental effort and performance. Paas et al. [58] describe mental load as load by task demand in relation to one's ability to deal with these. Mental effort is defined as the allocated capacities required for the task at hand. Hence, when cognitive load is measured, the effort is attempted to be determined. Additionally, to demonstrate the interaction and relation: when the task demand increases (mental load) the level of performance can be kept by using more capacities (increase effort).

The cognitive load theory (CLT) [71] distinguishes between three types of load: intrinsic load (ICL), extraneous load (ECL) and germane load (GCL). ICL is the load imposed by the intrinsic nature of the task and material. It depends on the user's knowledge and cannot be changed by design. ECL is the load imposed by the representation of the information of the task and material. The design of the material can be changed in order to reduce this unnecessary load. GCL is the load of putting effort into building new schemas. It depends on the ICL and, in combination, they build the base for schema acquisition.

Cognitive load is a term often used in HCI literature, but often it is not clear which of the three types is measured. The most common goal is to reduce the amount of extraneous load as it can be changed by design. Mental effort and load can be seen as the factors that affect the state of the working memory; hence, when attempting to measure the state, changes might have been caused by these two constructs.

2.1.3 Summary

The working memory builds the basis for several psychological theories, so it does for this work. The term cognitive state was chosen as a reference to the state of the working memory. The constructs, such as mental effort, might explain state changes. The next important aspect is to understand how the state can be measured.

2.2 Measuring the Cognitive State

There are several methods to assess the cognitive state such as subjective, objective, direct and indirect measures [46]. For instance, the NASA-TLX [33] is a questionnaire to measure workload, hence, a subjective and indirect measure. In this section an objective and direct measure is discussed: physiological measures. A short introduction to psychophysiology is given, followed by a description of the different physiological activities that can be used to assess the cognitive state.

2.2.1 Psychophysiology

Psychophysiology describes the research that investigates relations between psychological constructs and physiological reactions. These reactions are triggered by the central nervous system (CNS) which reacts to sensory or motor impulses very rapidly [60]. The human body has several systems such as the cardiovascular system, all of them work together to keep the body functioning[1]. The nervous system interacts with all systems and is, therefore, taking over a crucial function to keep so-called homoeostasis [60].

In physiological computing one of the important challenges is the psychophysiological inference (PPI) [24] which describes finding a valid mapping between the psychological constructs and bodily reactions or responses. Finding valid mappings is, however, not trivial. There can be a 1-to-1 mapping, many-to-1, 1-to-many or many-to-many. Ideally, researchers would like to obtain a 1-to-1 mapping, which would be the case for example if a change of the pupil size was only related and

[1] https://en.wikipedia.org/wiki/List_of_systems_of_the_human_body (Accessed: July 10, 2017)

caused by a change of the working memory state. In general, this is not true as light and other factors also influence the size. Figure 2 gives an overview of the different mappings and of how they can be described regarding their specificity and generality.

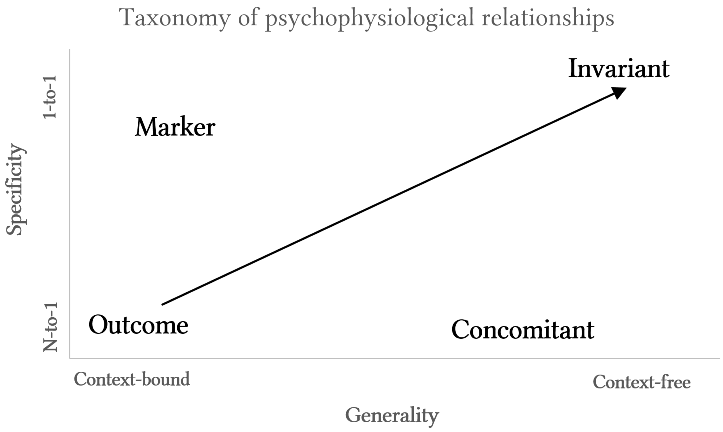


Fig. 2. Taxonomy of psychophysiological relationships. Adopted from [15]

Further, four types of relationships between the constructs and responses are shown. First, the marker relationship which isolates one construct and which can find a relation to one single measure in a very controlled setting. These kind of relationships are the ones commonly found in the literature. A 1-to-1 mapping in any setting (context-free) is described as an invariant relationship: it is the one that is desired to be found as it would allow using the findings in the real world. If more than one response changes in the laboratory it is called an outcome and outside of the laboratory a concomitant relationship. In the figure depicted the many-to-many case is not present as having the such is not helpful. Knowing that many responses changed when many psychological constructs where modulated won't answer any questions.

For this work, the existence of already found mappings such as mental effort affects pupil size are used to decide which measures can be used to assess the cognitive state. This is discussed in the next section.

2.2.2 Psychophysiological Measures

To get a hold of the bodily reactions physiological measures can be used [53]. In the previous section, it was mentioned that the body consists of different systems such as the cardiovascular system. In principle, the different systems can be measured or rather their responses. Cacioppo et al. [15] thoroughly discuss the possibilities. In this subsection first, the used measures in the experiment are introduced. The

reasoning behind utilizing these specific measures - pupil, skin, heart - is based on a previous analysis regarding their suitability to assess the state in near real-time and in an unobtrusive way [73]. Pupil measures were chosen due to the rapid response of the pupillary system, making state changes visible. Further, eye-trackers, used for measuring, are a rather unobtrusive technique in contrast to electrodes placed on different parts of the body. More detailed reasoning can be found in previous work [73]. Individuals differ in their responses, therefore using multiple measures can be useful. As a result, skin conductivity and heart-related measures were chosen to be used additionally. While they both can be measured in an unobtrusive way their suitability for a real-time approach required for adaptive systems is a challenge.
As literature discussed in this work also makes use of other measures an overview of other possible responses that have not been subject of the experiment is given.

Pupil. The pupillary system is described in [8] in detail, and most of this paragraph's content is based on it if not stated otherwise.

Pupil measures attempt to asses the pupillary response which controls the pupil dilation. The diameter can range from 1mm to 8mm [49] (BNID: 105349). When investigating the relation of the pupil to the cognitive state one needs to be aware of the pupillary light reflex which controls the amount of light let through - the bigger the pupil, the more light will get in. So in a dark room, the pupil is dilated to let more light through compared to a room filled with light. The dilation caused by light is much bigger in contrast to changes caused by the cognitive state. There is research that tries to account for light effects when using the pupil as a measure ([62], [76]). Further, while the term dilation has now been used several times, there is also constriction. Different muscles control both processes, and hence, their occurrence is interpreted differently. Constriction usually happens while the eye is accommodating and as a reaction to light while behavioural and stress contexts are associated with the dilation [9].

As stated by Beatty and Lucero-Wagoner [8] the diameter increases in task conditions with increased difficulty, and hence, can be used to identify variations of task demands. In [18] a memory recall task was used. They found a relation between cognitive load and changes of the pupil size. Another study [36] found indications that different levels of mental workload, modulated through a document editing task along with a route planning task, affected the percentage change of the pupil size. There is more literature investigating the relation between the cognitive state and the pupil size; however, there also exist relations to emotions. For instance, Bradley et al. [13] showed affective pictures to their participants to investigate the relation between arousal and the pupil. They found strong support for this relation. In another study [59] arousing sounds were used and indications for the relation of arousal and pupil size as well were found. The challenge of psychophysiological inference becomes very apparent in this case as changes of the pupil are related to different constructs.

Actual measures of the pupil are usually statistical measures such as the mean over a certain period. Other attempts are to use frequency analysis of the pupil signal which became known through the *Index of Cognitive Activity* [45]. Unfortunately, this index is not freely available. Tackling this issue is a recent paper by Duchowski et al. [22] who provide Python algorithms for an openly available version. To which extent it can be used, however, needs more investigation.
The most common hardware to record pupil and in general gaze data are eye-trackers. They can either be stationary (e.g. a webcam) or mobile (e.g. glasses). Usually, an image of the eye is used to detect the pupil and its size. A benefit of eye tracking is its unobtrusiveness for the user.

There exist other eye activity such as endogenous blinks, saccades and fixations [35] that also have been related to the cognitive state (e.g. in [17]). These measures, however, are very task dependent. For instance, in a visual search task, the user will have more saccades when trying to find something in contrast to a task where the user is asked to concentrate on one spot on the screen.

The pupil is a widespread measure when investigating the cognitive state, it changes rapidly, and the change can be unobtrusively grasped with eye-tracking technology.

Electrodermal Activity (EDA). This paragraph refers to [21] if not stated otherwise.

The electrical activity of the skin is often described with the term galvanic skin response (GSR) or electrodermal activity (EDA). The signal is usually measured with electrodes placed at the sweat glands of the fingertips, palms or feet. There are a tonic and a phasic component in the signal. The former is referred to as skin conductance level (SCL) and the measure of electrical conductivity over time. In contrast, the phasic component is referred to as skin conductance response (SCR). A response is a peak in the conductivity signal caused by a certain event or stimulus. There also exist so-called non-specific responses (NS-SCR) that are not triggered by an event such as a cognitive state change, hence, when utilizing SCR these responses should be filtered out which is not trivial.

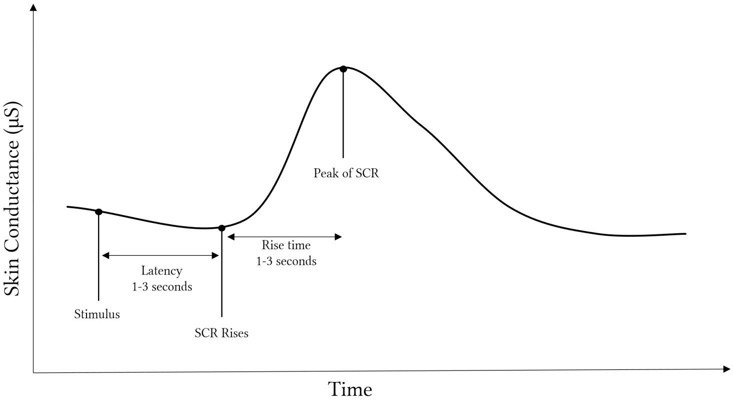


Fig. 3. Conceptual EDA signal. Adopted from [21]

Figure 3 shows a conceptual signal of skin conductivity. It can be seen that a latency issue comes with the signal. When a stimulus triggers a response, it takes 1-3 seconds until the signal starts to raise and another 1-3 seconds until the peak is reached. The response can be perceived best when the signal is measured at the sweat glands of the hands or fingers.

Dawson et al. [21] state that EDA can be used for measuring different processes: "activation, attention, and significance or affective intensity of a stimulus". Not only the cognitive state can be measured with it but also, and more prominently, arousal. For instance, increases in the number of SCRs have been found to be related to increased arousal in an experiment by Lang et al. [38]. In which participants had to look at pictures of varying valence and arousal. Further, Dawson et al. [21] summarized the use of SCL for arousal and alertness states while SCR might be more useful for attentional processes. Relations to the cognitive state have been investigated by Shi et al. [69] who found a relation of increased mean EDA and increased cognitive load. Two studies ([56], [55]) used accumulated EDA. In the first significant differences in accumulated EDA were found for eight arithmetic tasks varying in difficulty (four levels). The follow-up study used three more measures: frequency band power of EDA, blink number and rate. These measures were used to train machine learning classifiers. It was found that all four measures separately were able to predict the cognitive load level moderately well and that the accuracy can be increased by combining blink and EDA features.

EDA can be a cheap technique to get an indication for the cognitive state. It is quite unobtrusive and harmless for users. Due to its multiple causes of activation, it might be challenging to discriminate between emotional and stress-induced responses and cognitive changes.

Cardiovascular Activity. Physiological activity resulting from the cardiovascular system's functioning has several measures that were subject of psychophysiological investigations. One measure is the heart rate (HR) that refers to the number of times the heart contracts to move blood through the body. Similar to the HR is the pulse. In principle, the latter describes the same as the HR but to be precise, it is the number of times artery contractions occur which are caused by heartbeats. The variation of intervals between consecutive heart beats is a measure referred to as heart rate variability (HRV) [43]. The electrocardiogram (ECG) for which electrodes have to be placed on the skin near the heart, at the arms or legs [44] is an obtrusive method to get a hold of HR and HRV. For a less obtrusive method, today wearables make use of photoplethysmography (PPG) to measure blood volume pulse (BVP). The latter is derived from the blood volume (BV) captured with PPG using a photoelectric sensor [10]. To which extent wearables can be sensitive to psychological changes is an area that needs further investigation.

Literature's focus lies on the HRV measures [37]. Possible measures to derive from the HRV are thoroughly discussed by Malik [43] who proposes measures suitable for short-term HRV analysis. He states that two minutes of data are required for the analysis in a medical context. Cowley et al. [20] state that shorter time intervals ranging around 30 seconds might be sufficient. An attempt to address this issue was made by Zhou et al. [80] who investigated whether the raw BVP signal can be utilized for short time periods. After preprocessing the signal smoothing and standardization with the z-score different statistical features (e.g. mean), peaks (mean of peaks), maximum amplitude and frequency features were analysed . Significant results were found for the peak and maximum features. They computed these features for a whole task time that ranged from 20 to 120 seconds.

The relation to the cognitive state is seen as plausible by Berntson et al. [10] as different studies did show that decreases in HRV are related to increases in mental effort and workload. Rowe et al. [67] showed that HRV could be used in different ways. First, they showed that decreased HRV relates to increased load. Second, in the state where capacities of participants exceeded, HRV increased which is assumed to happen due to disengagement as a result of overload.
HRV is also used for emotion recognition ([54], [72]), hence, changes could always be related to emotional changes rather than working memory changes.

Further, HRV is synchronized with respiration, during inhalation the interval between two heartbeats is shortened and longer during exhaling - the respiratory sinus arrhythmia (RSA) [10]. It is usually measured when the HRV signal is transformed into the frequency domain, the high-frequency band (0.15Hz - 0.4Hz) is generally also called the RSA.

Cardiovascular activity is traditionally measured using an electrocardiogram which is rather obtrusive and not suitable for an interactive system. With smart wearables alternatives are available which, however, might lack in their validity and reliability to estimate heart rate and its derivatives.

Other Measures. The most prominent other measure to assess the cognitive state is electroencephalography (EEG) that measures electrical currents produced by the brain's neurons [44]. The signal is usually processed into the frequency domain which is categorized into different brain waves. These are alpha [8-14 Hz], beta [14-30 Hz], gamma [30-50 Hz], delta [1-4 Hz] and and theta [4-8 Hz] waves [44]. Each wave can be related to different mental states [52]. For instance, alpha waves are usually related to a relaxed state while beta waves are related to a state where more cognitive capacity is used [66]. Therefore the latter could be used for measuring cognitive load [52]. An overview can be found in [2]. An alternative method to assess brain activity is the functional near-infrared spectroscopy (fNIR). It uses infrared light to detect changes of blood's oxygenation (haemoglobin in the red blood cells) on the scalp - it monitors the prefrontal cortex of the brain [4]. Examples of the relation to the cognitive state can be found in the worky by Ayaz et al. [4].

Grassmann et al. [29] analysed 54 studies to investigated the relation of respiratory activity to mental or cognitive load. Respiratory effort is measured that is the "movement of the action of breathing" [40] in contrast to the analysis of the exhaled gas. Typically measures under investigation are respiration rate (RR) and variability (RRV).

2.2.3 Summary

The different physiological responses that might be used have been presented in this section. For all of them, research exists regarding their relation to the cognitive state. Other factors, besides working memory state changes, that might trigger responses have been pointed out. These other factors, such as light affecting the pupil, are important to consider when conducting experiments with physiological data but also when interpreting the results of the measures.

The next section will discuss the literature attempting to detect users' cognitive state with the help of psychophysiological measures in the context of human-computer interaction.

2.3 Related Work

In this section discusses related work with a focus on research that made use of physiological measures to assess the cognitive state in a near real-time manner, are related to adaptation and can be seen as HCI research.

The related work is categorized and discussed regarding the task settings that have been used to detect the state for but also how these were used to change the cognitive state. The so-called n-back task has been used in several studies. It is a working memory task, where participants have to store a sequence of stimuli and compare the newest stimulus with the one seen or heard n-steps before. Section 2.4.2 contains a more detailed description. Other working memory tasks will not be explained in detail in this section. Then a glimpse at the general goal of being able to adapt to the user is taken along with the challenge to assess the cognitive state. This is followed by a discussion of the methodology applied to extract the cognitive state information from the measures - machine learning.

2.3.1 Dual-Task & Single-Task

The tasks and their environment used in the literature range from real-world driving tasks to very abstract working memory tasks. Nonetheless, most of the tasks were done in controlled settings.

A standard methodology when investigating the cognitive state is to use a dual-task setup. Where a primary task, usually similar to a real-world task, is done by the participants. They as well have to deal with a secondary task which is supposed to trigger a change of the cognitive state. By using this approach, the physiological data can be labelled. For instance, in [70] participants were driving with a car on the road and the periods where the secondary task in the form of an audio n-back task (explained later) was present, were considered to be states of "elevated workload". In [64] working memory span tasks presented as notifications, users had to respond to while driving in a simulation (ConTRe task [42]), were used as a secondary task. While the second task did not change during the primary task in the examples above, Elkomy et al. [23] used various working memory tasks as secondary tasks during a Lego assembly task. In the context of Sweller et al.'s [71] cognitive load theory, the secondary task approach is used to detect increased demand of the primary task by using performance measures of the secondary tasks which are expected to get worse with increased primary task demands. At the same time when the secondary task is present also decreases of performance can be expected in the primary task which can make this approach tricky.

Using a single task to investigate the cognitive state is as well common; however, in this case, the tasks are usually simple cognitive tasks that can easily be adjusted in difficulty with a predictable effect on performance. Such simple tasks are often used as the secondary task that has been mentioned already. The n-back task and variations of it were used in [31] to produce different cognitive states. In [81] participants had to solve arithmetic additions of four numbers. 12 different difficulty levels were defined which represent different cognitive load levels. Multiple single tasks can as well be used as done by Ferreira et al. [26] who used two elementary cognitive tasks that are common to measure "perceptual speed and visio-spatial cognitive processing capabilities". In [17] three working memory tasks (arithmetic task, visual search, spatial task) each with varying difficulty levels were used to modulate the cognitive state. They continuously switched between these tasks to investigate task transition.

Dual- or single-task difficulty is used in both approaches to increase the working memory load. Section 2.4 will focus more the modulation of cognitive state. Further, all of these studies used the tasks discussed with different goals in mind. Still, the concepts of working memory, cognitive load or mental workload were central as well as the use of physiological measures and the goal of assessing the user state. Next, the challenge of assessment in the context of adaptation is discussed.

2.3.2 Adaptation the Goal, Cognitive State Assessment the Challenge

The introduction of this thesis already stated that for being able to adapt to users' cognitive state, first, the detection of it needs to be addressed. Some of the literature attempts to address both aspects but as there is no established cognitive state detection all of them had to deal with this issue first before being able to adapt. In [64], [23], [81] and [78] adaptation of a system was the general goal. The method that was used for the assessment is supervised machine learning. Except for Elkomy et al. [23] who defined a threshold for their physiological measures which triggered adaptation if surpassed. For supervised machine learning, it is required to have training data to built predictive models. This results in the necessity of either using data from previous studies [64], have a particular data gathering study [81] or have a sufficiently long training period during the experiment of which the data can be used [78].

With a system to detect the state, it is possible to investigate the effects of adaptation. For instance, user performance during the driving simulation task increased [64] when adaptation was used. Two approaches to keep the user in neither a too low or too high level of cognitive load were investigated in [81]. Avoiding overload leading to increased performance was the subject of [78]. Lastly, how adaptation affects user acceptance was investigated by Elkomy et al. [23].

Furthermore, studies without adaptation investigate the challenge of this assessment in near real-time which can be seen as data gathering studies ([31], [26], [17], [70]). Those studies have in common the goal of gathering physiological data using different task paradigms during various cognitive states of the user which then can be used as input for supervised machine learning. As adaptation was the motivation but not the goal, these studies investigated the challenges of the assessment. Ferreira et al.'s [26] focus was to investigate how several measures might be used in real-time to assess the state of users with varying age and gender resulting in them being able to distinguish between low and high load independent of these variables. Similar, Grimes et al. [31] investigated one particular measure, EEG, as input for machine learning finding trade-offs between training data size, window sizes and classification accuracy. For the study in [70], the context of driving was important. Further, they used data from roughly 100 individuals to asses the state in combination with performance measures leading to promising classification results. Gevins et al. [28] addressed the challenge of being able to assess the state for different tasks for different users with neural networks and EEG. They were able to reach good predictability with their network. Baldwin and Penaranda [6] and Walter et al. [75] based their research on the findings of Gevins et al. [28].

A question one might ask when looking at these studies is: whether the cognitive state is detected or the task the user is in is detected. They, same as this work, base their research on the assumption that changes in physiology are related to changes in psychological states, and they assume that this state is intentionally changeable by using task difficulty and common cognitive tasks. Rajan et al. [64] describe their detection of the state as detecting whether the user is only performing the primary task or the primary and secondary simultaneously. In other words, the task the user is currently performing is detected. Similarly, Zhou et al. [81] mapped 12 difficulty levels to either two or four cognitive states, and based their detection on these difficulty levels. The validity of these studies to assess a particular cognitive state is thus questionable. Therefore, looking back at the psychophysiological inference and relationships (see figure 2) one needs to be aware that in HCI literature such relations are not necessarily found but expected to hold.

The general idea is to exploit already found psychophysiological relationships, for instance, increased pupil dilation is related to increased working memory load. In most experiments, multiple measures, and sensors to capture them have been used. It could not be expected that every physiological response is sensitive to the task setting. One of the measures usually stood out in its capability to assess the state. Rajan et al. [64] used eleven different measures among them eye-tracking, electrocardiogram (ECG), photoplethysmography, EDA, respiration and skin temperature. Only one measure proved to be sensitive to their dual-task setup, namely, the pupil dilation. Which as well counts for [17]. Further, Wilson and Russell [78] and Ferreira et al. [26] found EEG to be more sensitive than ECG and respiration, also GSR for [26]. GSR was most useful in [23] and [81]. In the latter it was the only measure. Finally, HR appeared to work better to assess the state compared to EDA measures [70].

All of these studies had different settings, tasks and participants, hence, the variance of a measure being sensitive to the task can be explained. Choosing an appropriate measure for the cognitive state assessment is not trivial as it can not be expected that every measure is sensitive to every task and setting.

Whether studies aim for adaptation or assessment of the state what they share is the approach to make use of supervised machine learning which is discussed next.

2.3.3 Machine Learning - The Assessment Tool

Before discussing the related work regarding the machine learning (ML) approach, a short introduction to the topic is given. In the context discussed here, machine learning refers to supervised ML which is an approach to learn from examples, thus, learning from the physiological data gathered in the studies. It was not investigated if unsupervised approaches exist for assessing the state. To learn from data it needs to be labelled, which in the present context requires to label the physiological data with cognitive states or rather with task conditions representing various states. Besides labelling other steps are necessary such as preprocessing the data (e.g. filtering), extracting relevant features from it (e.g. mean SCL) and train a predictive model with these features (e.g. a state vector machine (SVM)). Feature extraction is usually done using sliding windows. Those are several second long segments of the physiological data. A model can be called individual if it was only trained using data from one person or it can be called a population model if the data of all users have been used to train one single model. Further, a metric is used to tell whether a trained model performs well such as the accuracy. To evaluate the results of a classifier the data used for training is usually split into one part used for training and one for testing. In the literature referenced here often cross-validation is used to train and validate a model on different splits of the training data. The average performance metric of models trained with different splits gives a better picture of the performance of the model. Classification can be used to distinguish between different amounts of classes, for instance, a 2-class or 2-way classification might discriminate between low and high load, a 3-class classification between low, medium and high. After this brief introduction to some terminology of machine learning, the related work is discussed.

Individual and Population Models. A trade-off between individual and population approaches might be the amount of data required per individual. For instance, Solovey et al. [70] used about 40 minutes of training data per participant when training individual models and were able to reduce the amount of data to four minutes when training across users. Further, in [81] a short practice phase was sufficient for each participant to calibrate their population model. It was trained on data of 10 participants for an online classification approach. Looking only at individual models in [31] only a marginal increase in performance was achieved by adding more data, they compared five minutes to up to 40 minutes of data. Ferreira et al. [26] suggest that only a short training period is necessary even for individual models.

As physiology differs among individuals models are required to account for that. Individual feature selection might be utilized to account for these differences ([78], [26]), however, to which extent this can be transferred to population models was not yet investigated. For the latter, the approach by Zhou et al. [81] mentioned earlier or a hybrid approach, individualizing the population models for each new user, as suggested in [70], might be used.

It seems that despite between-subject differences training across users is resulting in good performance, e.g. [28]. Rajan et al. [64] had similar accuracies for both individual and population models and the latter when used in their adaptation system with a new user resulted in increased task performance. For their 2-class classification Zhou et al. [81] reached up to 89% accuracy. However, this was achieved for an artificial task setting that makes it hard to generalize. In contrast to these rather promising results is the unsuccessful attempt to classify across users in [31]. Nonetheless, the focus there was on individual models. Appel et al. [3] proposed an algorithm for cross-user classification where for each participant a model is trained. For a new user, a short training phase is used to find similar subjects regarding the physiological response. Thus, every individual model already trained is assigned with a similarity score. It can be considered as a population model approach, even though they do not train a single model with data from all individuals but instead, use the individual models, similarity estimation and a voting scheme. They reached accuracies over 80% based on data of the n-back task.

Thinking of an interactive system the least amount of setup time is desired. If it is possible to reduce this time by using population models then they should be explored more. A question to ask is how a final model can be evaluated beyond using the same task and setting, when it is in use during interaction and when there is no ground truth to which the classifiers' output can be compared. For the same task, of course, this is possible but as soon as the model is supposed to be used as a general cognitive state classifier the challenge of evaluation rises. For instance, in the driving simulation setting by Rajan et al. [64] data for training was gathered with a notification containing working memory span tasks as the secondary task. In their particular setting they wanted to mediate notifications, thus, in order to use their classifier for that, they had to change the secondary task. Instead of validating the classifiers' performance they then used task performance measures to see if the adaptation has an effect. However, no general validation could be done to say whether the cognitive state was detected accurately.

Within-Task & Across Task Classification. The recent discussion raises the question of how not cross-user classification can be done but also cross-task classification. In Rajan et al.'s [64] case there are indications that changing the secondary task did not affect their classifier, but this is only an indication from looking at task performance measures. Besides heavily investigating the use of EEG Grimes et al. [31] let their participants perform the n-back task in different variants. They kept the structure of the task the same but changed the stimulus type (letters, images, spatial grids). Using only one type (letters) to train the models they were able to classify the other two variants with 80% and 77% accuracy. In contrast, Baldwin and Penaranda [6] were not able to achieve good performance for cross-task classification. However, their three working memory tasks were not as similar as in [31]. The same goes for Walter et al. [75] who attempted to classify the state for a diagram and algebra task based on data from three different working memory task. Their cross-classification did not perform well. The basis for the above research was built by Gevins et al. [28] who used data from two variations of the n-back task with three difficulty levels to classify across tasks. They achieved high accuracies for individual and population models. In the same way, one might contrast individual against population models it can be asked whether training for each new task is desired in contrast to having a model working across all tasks. Especially in a learning context where the task is new for the user and gathering of training data is not possible more general models are necessary [75]. While the latter is, for now, science fiction small steps might be taken as suggested by Ferreira et al. [26] who state that it might be possible to find a real-world task consisting of similar subtasks comparable to their elementary cognitive tasks, used in their study, for which their model can be used.

Towards Real-Time Classification. Classification might be done for a single user or across, might be attempted for a specific task or across tasks. If adaptation is the goal, near real-time detection of the state, is required. With trained machine learning models classification is a matter of milliseconds but before having appropriate input physiological data needs to be processed in real-time. Appel et al. [3] mentioned processing time as a general issue but had no troubles with processing their data in time. Part of the real-time approach is to detect the state based on small fragments of the physiological data using (sliding) windows. Grimes et al. [31] report a trade-off between window size and classification accuracy - the bigger the window, the better the performance. Their sizes varied from two seconds to 120 seconds which was a complete task sequence. Up to 30 seconds were used in [70] who report the same trade-off. In [31] the curve was quite steep starting with small window sizes and plateauing at the end with large windows. Both of these used different physiological measures: EEG [31], HR and EDA [70] showing that this trade-off might exist for different measures. The window size working best was seven seconds in [64] using pupil measures which, however, was their biggest window size. Similarly, five seconds were sufficient in [78]. In [26] once more the trade-off was pointed out as they compared 10 and 60-second windows. Depending on the physiological measure used the information of the cognitive state might be more visible when using larger windows in contrast to smaller ones. However, short-term changes cannot be detected any more when a large window is used. Especially in the case of the pupil dilation used in [64] small windows make sense as the pupil's processing is not time-consuming. Thinking of EEG where possibly artefacts have to be first detected (e.g. electrical activity produced by muscle activity rather than firing neurons) and then removed, frequency domain transformations have to be done or other data processing steps, more time might be needed.

What else drives classification performance. Window sizes, training data, individual vs. population and more have already been discussed as factors influencing performance of the machine learning models. The two last aspects discussed now are the features and the algorithms. For classification different algorithms have been used in the discussed literature such as Artificial Neural Networks (ANN) [78], [6], [28]; Random Forest Classifiers (RFC) [64], [81], [3] or single Decision Trees (DT) [70]; Support Vector Machines (SVM) [81], [75]; Naïve Bayes [31], [70], [81]; Logistic Regression [70], [17]; Multilayer Perceptron (MP) [70] and 1-Nearest Neighbour [70]; Quadratic Discriminant Analysis (QDA) [26]. A clear best choice algorithm, however, is hard to determine due to the different settings and physiological measures used. Further, as stated by Solovey et al. [70]:

> "[..] the classifier choice did not make a large difference in the results, showing that feature generation and selection are key to accuracy in this domain".

While the choice of a specific algorithm might not be that vital, the literature indicates that the choice of features can increase performance for individual models [70], [78]. Features are extracted for each window, thus, on a segment of the data such as 10 seconds of data. In [31] the effect of the number of EEG channels on accuracy was investigated showing a rapid increase when using more than two. In the case of EEG for each channel, features can be computed. Wilson and Russell [78] found that using only the best features for the particular individual leads to better classification performance. There were about 40 different EEG features and three others (inter-beat, inter-blink, respiration interval). Physiological differences exist between subjects; thus, it is reasonable to assume that some features are more discriminative for one user than for the other. It remains an open question how a per participant feature selection can be applied when using population models. For other physiological responses often the common and known derivatives are used. Such as skin conductance level (SCL) [81], [70], [26]; median of the pupil diameter in [3], [64]; heart rate (HR) [70] and variability (HRV) [32], [26]. In contrast to frequency domain measures used for EEG often statistical features are generated for pupil, heart and skin measures.

Besides physiological input other information can be fed to the classifiers as thoroughly discussed in [16] where also behavioural measures are explored (e.g. speech). In [70] driving velocity as a measure of performance was used as additional input which, however, did not add much to the accuracy.

Features used are not specific for this thesis topic as they are simply common when using psychophysiological measures. Most derivative measures for pupil, heart and skin have been introduced in the previous section. The literature discussed here did not perform any kind of feature generation method except for using statistical features.

Specificity of Classification. In the cognitive state section 2.1 cognitive load theory (CLT) distinguishing between three different types of the cognitive load was introduced. The literature discussed here; however, does not focus on discriminating between types but between levels of load. How many levels can be discriminated depends on the experiment setting and in most cases the number of different difficulty levels. In [70] using a dual-task setup, an "elevated level" of workload was classified when the secondary task was presented similar to [64]. These two examples distinguished between a normal level and an increased level. In single-task setups different difficulty levels represent different states of the working memory. An attempt to distinguish between four levels was done in [31] with accuracies less than 50% and in [81] with up to 64%. Better accuracies are reached when the classification problem is reduced to two classes, thus, two levels of cognitive state. For instance, in [75] two levels of the n-back task were classified correctly with an accuracy of 95%. While it might be advantageous to distinguish the three types of CLT the literature discussed here focused on merely discriminating between different states usually using task difficulty or presence/absence of the secondary task.

Following Baddeley's [5] theory where different types of information take up different capacities, it might be asked if each of these can be measured separately. However, it is not clear whether the types of information cause significantly different physiological responses. Gevins et al. [28] used the n-back task using letters with and without an extra spatial component, trained models with one task type and tested with the other reaching accuracies of 94% indicating that in this case, the different types of information did not affect the physiology differently. Similar attempts have already been discussed in the cross-task classification discussion.

Depending on the use case a distinction between high load and regular load might be sufficient as the driving example shows [70]. A distinction between two levels performs better than more.

Implications for using machine learning. From the discussion in this section, there were several factors identified relevant for cognitive state assessment. Models might be built across users or for each independently, which leads to trade-offs regarding generalizability, training data size and the accounting for individual differences. In some scenarios, a cross-task classification, where the models are trained with data from one task but used to predict the state of another task, is desired. For instance, if an entirely new task the user has never seen before wants to be evaluated.

The most prominent measure is EEG which did show promising results. Classification performance is driven by the choice of window size - the amount of data to create a learning example, and the amount used for classification. Further, the choice of features is relevant and can especially improve performance in the case of individual models as it allows to account better for subject-to-subject differences. The choice of algorithm is not the main factor resulting in a good performance. Concerning specificity, one can learn that discrimination between two states representing low cognitive load and an elevated load is more successful than making more fine-grained classifications.

The literature discussed here successfully made use of machine learning to classify cognitive state for different tasks, levels and users. Alternatively, other approaches using thresholds could be investigated further, e.g. [23]. If determining thresholds is automated the approach might be suitable but without that a procedure it is required to define thresholds manually. With machine learning, HCI researchers have an almost out-of-the-box tool for simplistic approaches to make classifications of the cognitive state. The space of possible approaches is not yet explored as expert knowledge is required to go beyond the rather simple approaches. The literature discussed here is mainly from the field of HCI and thus might have missed other relevant machine learning approaches which could be useful for cognitive state assessment. A supervised machine learning approach always requires ground truth data, thus, having a robust experimental setting is necessary.

2.3.4 Summary

The related work discussed used working memory tasks (sometimes as secondary task) with different levels of difficulty to artificially produce cognitively loaded user states that could be detected or predicted (e.g. [64]). By doing so, they were able to classify the state for a concrete task such as driving. This is hard to evaluate whether it might work in a real scenario where one does not know if the driver is cognitively loaded, and hence, classifier outputs cannot be compared with ground truth data. Other approaches not using such a particular task focused on basic cognitive tasks such as visual search or mathematical tasks (e.g. [17]). Here one has to ask to which extent generalization of these findings, using abstract tasks, is possible.

As an objective, near real-time measure psychophysiology can be used, yet related work highlights the typical challenges of it. Not every measure in every setting gives reliable data, for instance, only the pupil dilation responded well in the dual-task scenario in [64]. Every measure might be used, based on the literature pupil and EEG measures have proven their potential.

Keeping in mind that a cognitive state detection should happen in real-time, for instance, while driving a car, the method which can be used for it is machine learning. While training a model might take some time, predicting by using a trained model can happen in a matter of milliseconds. For the particular tasks used in literature, prediction showed promising results. Different factors relevant to machine learning have been discussed such as the features, algorithms or windowing. A challenge of predictive models is how they might be generalized so they can be used across individuals and tasks.

The focus and insights of related work were discussed in this section. Working memory tasks were omnipresent in the discussed literature the next section is going to continue with the question of how to modulate the cognitive state reliably to gather ground truth data for machine learning.

2.4 Modulating the Cognitive State

The related work revealed how the cognitive state might be modulated using either single-task or dual-task settings. The latter is suited for scenarios where a particular task is present, and an abstract task is used to raise the load. Another approach could be to have pre-defined difficulty levels for a concrete task where an easy difficulty level represents a state of low cognitive load; however, this requires an evaluation of the validity of these difficulty levels. One of the idea behind this thesis' work is to re-use the gathered data in one task for other tasks - inspired by [31], [28], [6], [75]. With this in mind, it appears to make sense to collect data during abstract tasks rather than focusing on a specific scenario like driving. To modulate the cognitive state working memory tasks (WMT) might be used. A benefit of these is that on the one hand there is much research, they are valid, and hence, they can be used to produce different states of cognitive load reliably. Additionally, these tasks focus on the working memory, and there is no need to choose a specific higher psychological construct discussed earlier - WMTs focus on the common ground of the psychological theories. Still, there are different models of the working memory as seen previously, for instance, Miller [48] and the +-7 chunks or the model by Baddeley [5].

In the following characteristics of working memory tasks along with some examples are presented. Then the chosen task for the experiment in this work is explained - the n-back task.

2.4.1 Working Memory Tasks

This section is based on the work by Conway et al. [19] if not stated otherwise, some formulations are taken from [74].

In [19] they review tasks that are typically used to measure working memory capacity (WMC). In contrast to measuring in-use capacity, the goal of measuring cognitive state in the thesis' context, measuring WMC refers to the overall capacity of an individual. For instance, different span tasks exist: reading, counting and operation span. The latter might work as follows: first, a letter is presented that needs to be remembered, followed by an arithmetic expression that the participant has to evaluate whether it is valid or not. Then another letter to memorize can be presented. At the end of a task sequence, all of the letters have to be recalled. How can this be used to measure WMC? In the first run of the task, there will be, for instance, two letters to remember, if the individual successfully manages to recall all of them correctly in the next run three letters are required to be memorized. This can go on and on until the individual is not able to recall the required amount of letters any more.

Physiology differs among individuals so does the WMC. Further, in the above example arithmetic tasks were used this could also be a sentence that is either true or false. One individual might be better at one or the other. Thus, Conway et al. [19] suggest to consider using different span tasks to account for these differences. However, the review does not make a distinction between verbal and spatial related working memory capacities or processing.

Task Name	Aspects	Description
Memory Span	Storing Visuospatial, Verbal	A task where elements of one or more item types have to be memorised and then recalled. Recall can be reversed or in order
Matrix Span	Storing Visuospatial	A task where a grid is displayed with one highlighted grid element, then another grid with another element and so on. At the end the positions of the highlighted grid elements have to be recalled. It is the visual version of a digit span task basically.
Visual Search	Processing Visuospatial	A specific object has to be found, there are multiple items serving as distractor that can vary in amount.
n-back	Storing, Processing Verbal	A sequence of items is displayed, the participant has to indicate whether the current presented element matches the one n steps back.
Dual n-back	Storing, Processing Visuospatial, Verbal	Two stimuli are presented, visual and via audio, for both it has to be decided if they match the one n steps back.
Spatial n-back	Storing, Processing Visuospatial	Instead of a sequence, an element in a grid is highlighted, and then the one n steps back needs to be kept in mind and matched against the current stimulus.

Table 2. Examples of working memory tasks

Working memory tasks should involve storing and processing of information. In the example, both aspects are separated, memorizing letters requires storing while the arithmetic task needs processing. In a continuous task, these two functions might overlap. Some examples of typically used tasks are presented in table 2 they are taken from different sources ([19], [51]) some examples also can be found and tried out on *James Stone's Cognitive Tools Page* [1]. Further, the tasks are categorized based on Baddeley's [5] theory to indicate which aspects the tasks are addressing mostly. It has to be noted that these aspects can be changed for most of the tasks, depending on the type of stimulus and presentation modality (e.g. visual or via audio). Thus there exist different variations of these tasks.

All of these tasks can be used to modulate the cognitive state. In this work, it is made use of the n-back task. As pointed out by Conway et al. [19] it is the *gold standard* task and it is used frequently in neuroscience studies to investigate the cognitive state. Further, it is a continuous task where storing and processing overlaps. Such a task resembles the situation where a user interacts with an adaptive system better. There the user state has to be detected continuously. The next section will discuss the task further.

[1] http://www.cognitivetools.uk/cognition/ (Accessed: November 21, 2017)

2.4.2 The n-back Task

In all variations of the n-back task, stimuli are presented and have to be compared with stimulus n-steps before. An example is given in figure 4 with a variant used in the two experiments of this thesis.

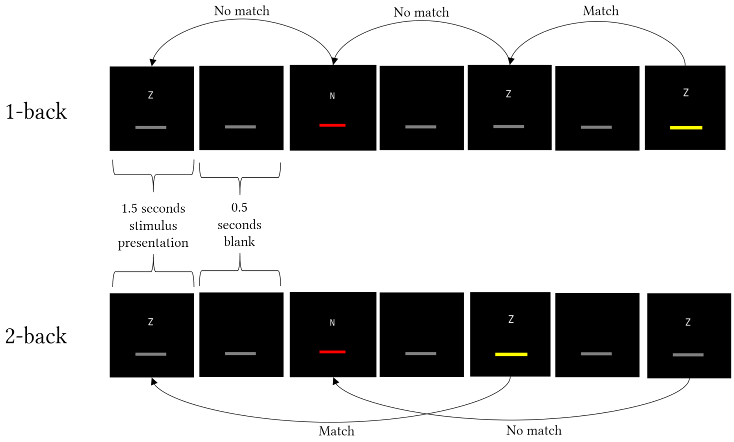


Fig. 4. 1-back (top) and 2-back (bottom) example of the n-back task. The grey, red and yellow bar are visual feedback to the participant's response (none, wrong, correct)

As indicated in the figure the n-back task has several adjustable parameters. First, the n is supposed to increase the difficulty and the common choice to modulate difficulty. Second, the presentation time of a stimulus can be varied, for instance, a longer presentation might help the user. However, it is not desired that there is enough time to rehearse. Third, the number of stimuli in total can affect the performance and therefore difficulty, e.g. fatigue effects. The stimuli can be presented either via audio or on a screen, for instance, letters or numbers can be displayed (verbal), a grid with highlighted elements can be displayed (visuospatial) or pictures of objects could be used. For the latter, an audio version might not be straightforward to choose. Performance might be measured using the response time or accuracy (number of correct answers).

In literature a clear definition of stimuli presentation time amount and break is not given. Related work used different variants of the n-back task some used a presentation time of 500 milliseconds (ms) followed by a blank screen or fixation cross for 1500 ms [3], [75] or 2000 ms [6]. Grimes et al. [31] used a presentation time of 1000 ms and a blank screen of 3000 ms. The shortest presentation time found was 200ms presented in intervals of 4.5 seconds [28]. In an n-back audio version using digits (also referred to as delayed digit recall) intervals of 2.25 seconds were used

[70], [27]. One n-back sequence ranged between 30 and 120 seconds; thus, a clear view on how many stimuli should be presented is as well not given. Audio was already mentioned as a modality; the others used visual and verbal information as stimulus. For instance, letters in [31] and [28] but also spatial information stimuli in the same two studies. Finally, the difficulty in the discussed studies did never go beyond 3-back, indicating that this might be sufficiently challenging to produce a high load.

2.4.3 Summary

Working memory tasks can be used to modulate the cognitive state of the user by increasing the information that needs to be stored and processed in the working memory. Different tasks require to deal with different types of information, for instance, spatial, verbal or visual information. Many tasks are not continuous in their original form. Therefore, those which are, are highly interesting as they resemble the real world better. As a result, the n-back task was chosen, and several examples from the literature were used to determine suitable task parameters.

2.5 Summary

This section introduced the relevant theoretical foundations this thesis builds on. The cognitive state is considered as the state of the working memory, and psychophysiological measures can be used to assess this state. Related work revealed several aspects important for pursuing the goal of detecting the cognitive state such as the choice of a task or the potential of machine learning. To dig deeper into the task topic working memory tasks were introduced.

While the broad research direction was discussed in the introduction, the next section will introduce the research questions this work attempts to answer based on the discussed background information of this section which informed the formulation of the questions.

3 RESEARCH QUESTION

In the previous section, it became apparent that physiological data in combination with machine learning can be used to classify the cognitive state of users for specific tasks, but also for rather abstract tasks. The data gathered for machine learning is very specific for the particular experiments, and it remains a challenge to re-use this data for training in other scenarios. There have been several investigations either tackling the challenge of cross-user or cross-task classification by making use of EEG measures with mixed results. On the one hand, there are promising attempts such as in [28] where electroencephalography (EEG) was successfully used for models trained for individuals and across all using two variants of the n-back task. They also were able to classify across tasks for individuals. Furthermore, cross-task classification with individual models yielded promising results in [31] who used EEG and variants of the n-back task. The pupil diameter was successfully used in [3] across users. On the other hand cross-task classification did not work very well in [6] and [75] both using EEG and working memory tasks.

To the author's knowledge, there has not been an approach of cross-task and -user classification utilizing pupil dilation measures in combination with heart rate and skin measures. Thus, the general goal of this work is formulated as follows:

Goal. Predict the cognitive state of the user, utilizing physiological data gathered during an abstract working memory task, using machine learning models trained with data of all users for different kind of tasks.

As a first step, an experiment with the n-back task using letters as stimuli is conducted during which physiological data - pupil dilation, cardiovascular and electrodermal activity - is recorded during different difficulty levels of the task. With this experiment, the first research question is addressed which is formulated as:

Research Question 1 (RQ1). Can the physiological data of the n-back task with letters as stimuli be used to distinguish between different levels of cognitive state using predictive modelling across users?

Different levels of the cognitive state from low to high load are supposed to be represented by the different levels of difficulty. EEG is the most investigated physiological measure in literature. In this work, the focus lies on the less obtrusive measure of pupil dilation captured with eye-tracking glasses. Physiology differs among individuals, and furthermore, not every task setting might elicit a physiological response as seen in the discussion of related work. Thus, as a data triangulation measure, two additional unobtrusive measures are going to be subject of investigation, namely, electrodermal activity and blood volume pulse.

In addition to that, the continuous n-back task is used to modulate the cognitive state, the version in the experiment uses letters visually presented; hence, verbal and visual stimuli. If decent classification accuracies can be reached with this data, it can be attempted to continue addressing the main research goal by using a different task, gather new data, and finally try to predict the user state with the model of the first experiment for the new task. The objective of making cross-task classification is formulated as the second research question:

Research Question 2 (RQ2). If population models work reasonably well for the n-back task with letters, can it be used across variants of the n-back task?

A subsequent experiment is going to investigate RQ2 using variants of the n-back task with different types of stimuli similar to Grimes et al. [31]. In contrast to Grimes et al. [31] unobtrusive sensors are used and cross-user classification is investigated.

In order to be able to build population models example data is required for which a data gathering study was conducted - the subject of the next section.

4 DATA GATHERING STUDY

The goal of the study is to gather physiological data for the supervised machine learning approach with a task that is expected to modulate the cognitive state reliably. This allows addressing the first research question concerned with building population models to discriminate between different levels of the n-back task. The experiment is supposed to elicit different levels of cognitive state by using three difficulty levels of the task. It is expected that the physiological data collected is sensitive to the three levels so it can be used for machine learning. In a pretest and pilot study discussed later, the pupil radius appeared to be sensitive. Therefore, it is expected to be sensitive as well in the larger scale study. The first part of this section describes the experimental study while the second part presents and discusses the results.

4.1 Experiment

In the experiment, participants performed the n-back task with letters of the Latin alphabet in three different difficulty levels, namely, 1-back, 2-back and 3-back. During performing, they wore an eye tracker and the Empatica E4 wristband. The experiment lasted for one hour, and participants were compensated with 10€. Additionally, three Amazon vouchers (30€, 20€, 10€) were promised to the top three performers. This was intended to keep them engaged in the task during all conditions. The details of the experiment are described in the upcoming subsections.

4.1.1 Design & Variables

A within-subject design was used, so every participant performed the task in all conditions. The independent variable was the difficulty level (n = 1, 2, 3). The order was counterbalanced with the *Williams Design* [1] [77] which uses *latin squares*. The dependent measure was the cognitive state which is operationalized with the physiological data: pupil dilation, electrodermal activity (EDA) and blood volume pulse (BVP) measures. Further, the NASA-TLX [33] was used as a measure of the cognitive state.

The light conditions were controlled as the pupil dilation is affected by it. Movement artefacts were kept to a minimum for the EDA and BVP measures by telling the participants not to move the arm to which the sensor was attached. Further, the number of matches (letters equal to the one n-steps before) was always one-third similar to [31].

4.1.2 Procedure

Participants were welcomed. A chair was prepared where bags and jackets of the participants could be put for the duration of the experiment. They were asked to take a seat and read the information regarding the study after which they signed the

[1] http://statpages.info/latinsq.html (Accessed: February 2, 2018)

informed consent. A short interview was used to get demographic data, memory game experiences and their current cognitive handicap.

The participant chair was a typical office chair which had adjustable armrests, and the sitting height could be changed. Rests and sitting height were set so that the participant was able to perform the experiment comfortably. Then the E4 with lead extension wires was attached to the wrist of the non-dominant hand. The eye tracker was put on afterwards: the recording quality was checked as well as the position of the eyes which should be centred in the eye camera view. A three-point calibration was performed before the experiment finally started.

The task was explained on the screen with a subsequent practice phase during which participants could ask questions. The study conductor observed if the task was understood and performed correctly and gave hints if there were any misunderstandings. Each task difficulty level, 0-back, 1-back, 2-back and 3-back, was performed with 30 stimuli each during the training phase.

After the practice phase participants were told that it is essential to stay focused on the task, to give their best and that the study conductor will not observe the participant while doing the task. The three difficulty levels were then performed in a counter-balanced order. After each level, the NASA-TLX questionnaire was handed out to the participants. Each difficulty level had four trial runs. A trial run is depicted in figure 5. After completing all three difficulty levels, the sensors were taken off, and a final semi-structured interview was conducted. To motivate the participants prizes for the best performances were promised. As a result, the score was computed while the participants were still present. Afterwards, a compensation of 10 € was handed out to them.

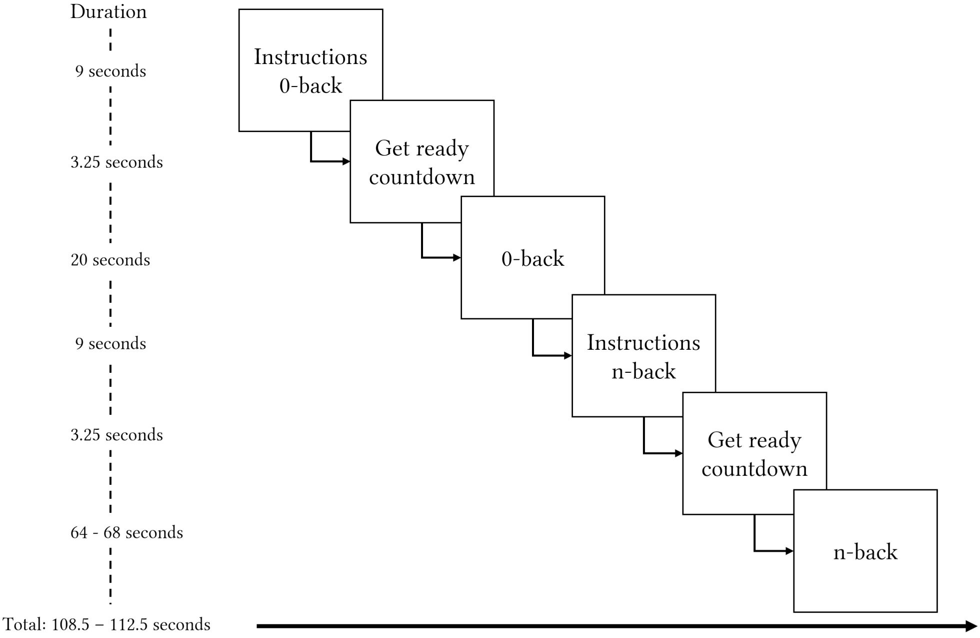


Fig. 5. A trial run: each difficulty level had four of these.

4.1.3 Task

For the n-back task, ten letters were chosen: (C, D, F, H, K, N, P, R, V, Z). They are highly readable [39], and the set does not contain any no vowels which would allow building small words [31]. The task was to tell - by hitting a key - for every presented letter whether it matches the one n-steps before. In the 0-back condition, a letter was presented during the instructions which had to be memorized and every presented letter had to be matched against it.

A small rectangular bar was used to give feedback on the response: it was coloured yellow if the response was correct and red if it was incorrect. In contrast to other work, feedback was included. During interaction with a real system, the user would receive feedback on its action which as well would indicate if an action is correct or wrong in the widest sense. In one single trial run (see figure 5) 30 + n stimuli were presented, so in every difficulty level, the participants had to respond to the same amount of stimuli. A single stimulus was presented for 1500 ms with an additional 500 ms where no stimulus was shown. A baseline difficulty 0-back preceded every trial run of the other difficulty levels. Here the number of stimuli was reduced to a list of ten letters with a random order. For the other difficulty levels the list of stimuli was generated (inspired by [31]) such that one-third were matches, one fourth were no matches but similar to the recent letters, and the rest were neither matches nor similar. Following additional rules were used: there were no three consecutive matches or similar, if there was no match in the last three steps it was forced, if there was no similar in the last four steps, it was forced.

The task was implemented using the tool *OpenSesame* [1] [47] which allows creating experiments with a user interface but also can be modified to one's needs with *Python* scripts.

4.1.4 Apparatus & Data Recording

The *SMI Eye Tracking Glasses 2 (ETG2)* [2] were used to record the pupil dilation at a rate of 120 Hz. The glasses come with a C/C++ SDK which was used to build a small application to be able to calibrate the eye tracker for every participant and record the gaze data.

The *Empatica E4 Rev2 Wristband* [3] was used to record blood volume pulse (@ 64Hz) and electrodermal activity (@ 4Hz). The wristband sends its signal via Bluetooth to a small server application provided by Empatica. Another application was written to connect to this server and receive the data for logging. Both of the small applications for the eye tracker and the E4 send their data to a central data collector application where the data is logged. It writes log files for every data stream with timestamps that allow for easy synchronization. For creating the timestamps, the *LabStreamingLayer (LSL)* [4] API was used. The task implementation also sends experiment state information to the central logger. A schematic overview is depicted in figure 6. The OpenSesame experiment ran, along with the sensor and data recording applications, on a single notebook - a Lenovo Thinkpad P50. The participant monitor was a 24" screen (HP LP2475w), the resolution was set to 1920x1200, and default brightness/luminance settings were used. A Logitech keyboard was used, it was turned vertically, and the numpad keys Enter, + and - were coloured in *blue, orange and white* and were used for *matches, no matches and to continue*. These were the only keys participants had to use during the whole experiment.

[1] http://osdoc.cogsci.nl/ (Accessed: January 13, 2018)
[2] https://www.smivision.com/ (Accessed: January 13, 2018)
[3] https://www.empatica.com/research/e4/ (Accessed: January 13, 2018)
[4] https://github.com/sccn/labstreaminglayer (Accessed: January 13, 2018)

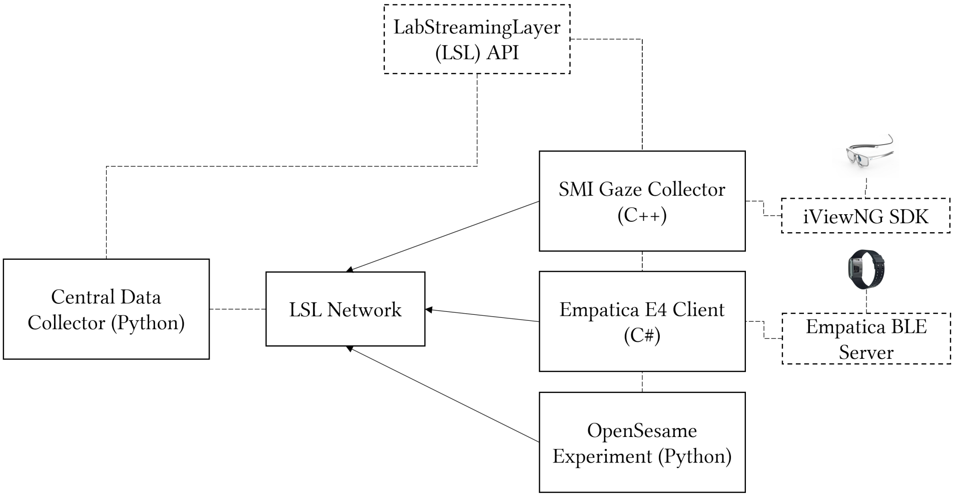

LabStreamingLayer (LSL) API
SMI Gaze Collector (C++)
iViewNG SDK
Central Data Collector (Python)
LSL Network
Empatica E4 Client (C#)
Empatica BLE Server
OpenSesame Experiment (Python)

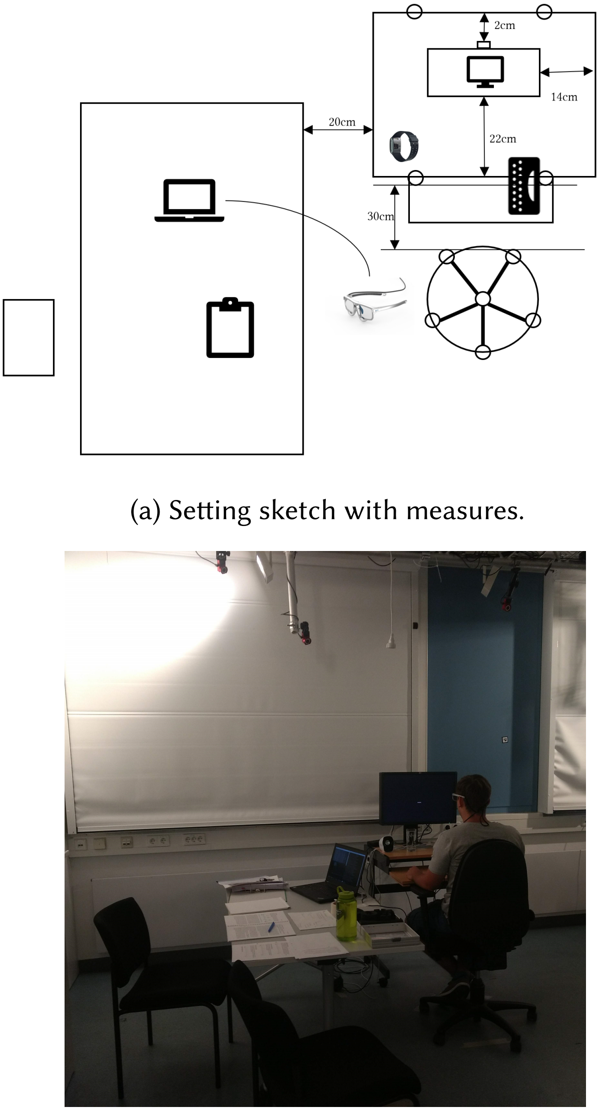
Fig. 6. Abstract view on the data collection architecture

Explanation

The LSL uses the network to communicate between different components. The sensor applications send their data on the network which are received by the central collector that is collecting data from all available network streams.

4.1.5 Setting

The setting is depicted in figure 7. In a) a schematic overview is shown while b) shows the actual lab setting. The experiment notebook was placed on a table left to the participant - the study conductor table. Next to it was the participant chair - an adjustable office chair - with the keyboard participants used to respond to the stimuli and the display on which they were presented.



(a) Setting sketch with measures.

(b) Picture of the real setting.

Fig. 7. Study Setting

4.1.6 Pretest & Pilot Study

The experiment described in this section was the final design that was informed by a pretest and a pilot study. The pretest was used to, first, determine whether the chosen task parameters modulate the cognitive state as expected: 1-back low load, 2-back medium-load, 3-back high load. Second, to get an indication that the pupil will respond to the experiment's task: is there a difference between the difficulty levels? Third, the pretest served as a technical evaluation of the study prototype and setting. Fourth, as the goal was to gather data for machine learning preprocessing and training models with the data was experimented with.

Four participants took part in the pretest, and the main outcomes of the analysis were the following. The task parameters were fine, so the three difficulty levels were perceived as expected. The NASA-TLX and subjective rating scales were used to measure the difficulty. The pupil responded well - there was a clear difference between the 0-back phases and the n-back phases but also between the difficulty levels of 1 and 2. A clear difference between levels 2 and 3 could, however, not be seen. Experimenting with machine learning as well resulted in promising results for distinguishing between levels 1 and 2. Several improvements were made for the prototype which enhanced the logging and task. For instance, the stimuli generation as described earlier was implemented.

After the pretest, and the adjustments made, before starting the study a pilot was run. Additionally, during the pretest, only the eye tracker was used while in the pilot study the E4 could be used to record data. The main findings were, the E4 did not record the electrodermal signal well, this was solved by attaching lead wires to the E4 that allowed to place the electrodes on the fingers. It increased the signal quality tremendously. The pilot did not reveal any new technical issues or necessary task improvements. Instead, it lowered the expectations for the signals coming from the E4 which did not show promising patterns as the pupil size did.

4.1.7 Summary

This section described the experiment in detail so other researchers can copy it easily. The next part presents its results.

4.2 Results & Analysis

The goal of the study was to gather usable physiological data of 24 participants. The order of the difficulty levels was counter-balanced leading to six different orders. Hence, six participants were necessary to fill one block of orders. To recruit participants, many flyers were put on the walls of the university campus. On the flyers, the study was advertised as *Gedächtnisspielstudie* which translates to memory game study. Additionally, some flyers were put on the tables of the university cafeteria.

The purpose of the analysis is to first make a treatment check by analysing the performance measures and the NASA-TLX scores. The treatment check will give insight into whether the three n-back task levels represent three different levels of difficulty and potentially different levels of cognitive load. The related work analysis has revealed the challenge that not every physiological measure might be sensitive to every experiment's manipulation; thus, before using the data as input for machine learning the analysis is supposed to reveal which measures make sense to use.

4.2.1 Analysis Procedure

To begin general remarks regarding the analysis approach are described. For each participant directly after participation, the data was inspected to check whether it could be used for analysis or if a replacement had to be found that would run the task with the same order. Six participants had to be replaced due to noisy physiological data or other reasons which will be discussed later. In total 30 participants took part in the study including the pilot study participant but only the data of 24 participants was used for the analysis reported here.

As a remark for the statistical analysis the following should be noted. Significant and positive results are coloured in green, those close to significance (positive or negative) are coloured in yellow and negative significant results are marked with the colour red. Positive significance can be found for an ANOVA or Friedman Test and their respective post hoc tests. While negative significance can be found in tests checking the assumptions such as normality thus, they indicate a violation of a required assumption.

Statistical tests are considered to be significant if their p-value is smaller than $\alpha = 0.05$ or even smaller as for the post hoc tests a Bonferroni correction is used. The experiment design is a repeated measures design with one independent variable - difficulty - with three levels. Either a repeated measures (RM) ANOVA with paired t-tests as post hoc tests is used if the data can be considered to be parametric. For non-parametric data, the Friedman test is used with the Wilcoxon signed rank test as post hoc test. Two assumptions of the RM ANOVA are sphericity - tested with Mauchly's test - and that the data does not significantly diverge from normality - tested with the Shapiro-Wilk test. In some cases, their results might not be reported directly, but in case the ANOVA is used it can be said that at least these assumptions were checked.

All tables regarding the analysis can be found in the appendix. While there the complete analysis is traceable in this section only the most important aspects are highlighted and reported.

The analysis of the data was done using *JASP* [1], *Python* using *NumPy* [2], *pandas* [3] and *SciPy* [4]. Plots, figures and tables were generated directly from JASP or with code using *matplotlib* [5].

4.2.2 Demographics

From the 24 participants used for analysis, there were ten male and 14 female participants from which all except two were students. The latter had finished their studies recently and were in the transition from university to the first job. The subjects were psychology (6), economy related subjects (4), physics (3), political science (2), linguistic-related subjects (2), law (1), life science (2), nanoscience (1) and English language and literature studies (1). The average age was 23.375 (SD = 3.146) years with a range from 19 to 29. Eight were wearing contact lenses during the experiment. None had particular experience with memory games except for the psychology students where some knew similar tasks, but they had no particular practice with them.

Besides these rather common questions, an additional question was posed: participants had to rate their cognitive handicap on a scale from 1 to 10. Rating 10 was described to them as a state where it is hard to concentrate (e.g. after learning for an exam for 8-10 hours). Rating 5 was described as a state where already demanding work was done but capacities, to at least get the same amount of work done, are still available. Rating 1 was described as a state where one feels fresh and ready to write an important exam. The idea behind this was that performance might be affected by the time of day and the stuff they did before the experiment. Linear correlation was tested with Pearson's r (see table 3) for the three variables, performance ($M = 0.878, SD = 0.04$), time and cognitive handicap ($M = 4.042, SD = 2.116$). The performance score was the average of the accuracy of the three difficulty levels. The time was converted into minutes (e.g. 08:15 to 495). The only correlation found was between time and the handicap rating which is not surprising. A linear correlation neither exists between the accuracy and the handicap rating nor the time and the accuracy. It cannot be said that the two factors have no influence.

[1] https://jasp-stats.org/ (Accessed: July 13, 2018)

[2] http://www.numpy.org/ (Accessed: March 4, 2018)

[3] https://pandas.pydata.org/ (Accessed: January 21, 2018)

[4] https://www.scipy.org/ (Accessed: January 13, 2018)

[5] https://matplotlib.org/ (Accessed: March 4, 2018)

Still, the correlation results indicate that. Another indicator for that is the quite good average performance of all participants.

			Pearson's r	p
Time	-	Cognitive Handicap	0.693	< .001
Time	-	Accuracy	-0.111	0.606
Cognitive Handicap	-	Accuracy	-0.132	0.537

Table 3. Correlation Matrix: performance as accuracy, cognitive handicap rating, starting time in minutes.

4.2.3 NASA-TLX, Performance

In order to be sure the difficulty levels of the task were perceived as expected, so that the assumptions of low cognitive load is represented by the easy difficulty (1-back), medium load by the medium difficulty (2-back), and high load by the hard difficulty (3-back), the NASA-TLX was used along with the performance measures.

The TLX measures workload using six dimensions: mental demand, physical demand, performance, effort and frustration. For each dimension, the participant gives a rating between low and high. The raw score of the TLX is computed by the formula :$\frac{mental\_demand+physical\_demand+performance+effort+frustration}{number\_of\_dimensions}$. It appears that some participants misunderstood the performance rating. In contrast to the other dimensions where the scale is labelled with low and high, for the performance the labels are good and bad. Some participants might have answered the question *How good was your performance?* with bad even if they thought their performance was good. This might have happened as they expected a high performance rating to represent a good performance. While it is not possible to reliably tell which participants mixed the ratings up it still needs to be mentioned and kept in mind when looking at the results of the questionnaire.

It was expected that the score for 1-back is smaller than the 2-back score and the latter less than the 3-back score. Table 4 shows the descriptive statistics of the TLX scores due to the task not being a physically demanding task the dimension was dropped. That, however, did not affect the general tendency. Hence, only the results with all six dimensions are reported below.

	TLX Raw			TLX Raw without physical demand		
	1-back	2-back	3-back	1-back	2-back	3-back
Mean	32.604	47.326	57.014	36.625	53.333	64.708
Std. Error of Mean	2.724	2.576	2.957	3.061	2.939	3.153
Median	31.667	50.417	58.333	34.500	57.500	65.500
Std. Deviation	13.343	12.619	14.488	14.995	14.397	15.448
Minimum	10.000	15.833	18.333	8.000	19.000	22.000
Maximum	63.333	65.000	91.667	66.000	76.000	91.000

Table 4. Descriptive statistics of the NASA-TLX ratings.

The workload ratings were as expected. A Friedman test was chosen as it does make fewer assumptions about the distribution of the data. The result ($\chi^2 = 28.667, p < 0.001$) indicates that the scores are significantly different. The Wilcoxon signed rank test was used as a post hoc test for pairwise comparison see table 5. As all were significant, it can be assumed that the three difficulty levels produced three different levels of workload.

Performance					
			W	p	Rank-Biserial Correlation
Accuracy 1-back	-	Accuracy 2-back	291.500	< .001	0.943
Accuracy 1-back	-	Accuracy 3-back	300.000	< .001	1.000
Accuracy 2-back	-	Accuracy 3-back	297.500	< .001	0.983
Response Time 1-back	-	Response Time 2-back	0.000	< .001	-1.000
Response Time 1-back	-	Response Time 3-back	0.000	< .001	-1.000
Response Time 2-back	-	Response Time 3-back	21.000	< .001	-0.860
NASA TLX					
1-back	-	2-back	26.000	0.001	-0.827
1-back	-	3-back	7.000	< .001	-0.953
2-back	-	3-back	24.000	< .001	-0.840

Table 5. Wilcoxon Post Hoc for Performance Measures and the NASA-TLX

For the performance, it was expected that the response accuracy and the response time is getting worse with increased difficulty. Table 6 shows the descriptive statistics of the performance scores. The accuracy was computed by the formula $\frac{number\_of\_correct\_answers}{number\_of\_possible\_answers}$.

	Response Accuracy			Response Time		
	1-back	2-back	3-back	1-back	2-back	3-back
Mean	0.968	0.896	0.770	583.554	739.276	836.635
Std. Error of Mean	0.005	0.012	0.015	16.717	23.290	23.540
Median	0.976	0.917	0.770	559.335	731.388	843.867
Std. Deviation	0.024	0.061	0.072	81.896	114.096	115.322
Minimum	0.887	0.742	0.621	470.790	563.677	621.387
Maximum	0.992	0.976	0.919	754.323	1016.266	1105.113

Table 6. Descriptive statistics of the performance scores (accuracy and response time).

As expected the accuracy was highest for the 1-back condition and lowest for the 3-back condition. The response time increased with increasing difficulty. A Friedman test was used for the performance data as well. The results ($\chi^2_{Accuracy} = 42.250, p < 0.001$, $\chi^2_{Responsetime} = 42.750, p < 0.001$) indicate that the scores are significantly different for each difficulty level. The Wilcoxon signed rank test was used as a post hoc test for pairwise comparison. For the accuracy all comparisons were significant (see table 5).

The performance results show that the difficulty was modulated successfully. In relation to each other, the three levels can be seen as easy, medium and hard.

With the results of the performance measures and the NASA-TLX, it can be said that the n-back task worked as expected, producing three different states. To which extent these three states represent a cognitive load level of low, medium and high, however, is not proven with this. It can, however, be assumed that all three levels elicit different states with each being different in cognitive demand.

4.2.4 Post Interviews

The semi-structured interview was intended to get an idea of how the task was approached by the participants, if they had to guess their answers and how the difficulty jumps were perceived. The interview was quite open, and in the following, some interesting remarks by the participants are highlighted. Some of the answers that were similar are reported with a frequency.

Guessing. 20 participants reported that they had to guess in the 3-back condition. Ten of those said that it happened rarely and two said it happened in more than 40% of the time. One participant stated that the letter presented at the beginning of the 0-back condition was forgotten due to not being concentrated, and hence, guessing was necessary. 14 further explained that they were guessing intuitively. That means, for the current letter they were trying to figure out if it was not similar than any of the recent letters. If it was similar a probability existed that the letter is a match. The guessing was therefore not blindly pressing the blue or orange key. Guessing followed after making a mistake which lessened the confidence that they had remembered the sequence correctly.

Difficulty. When asked if any of the conditions was to difficult nine answered with 3-back might be too hard. For instance, P34 said that more time is required to memorize the sequence which made it too hard. Further, P39 stated that there is a certain amount of time necessary to develop an approach to deal with the 3-back difficulty. In contrast, only four stated that 1-back was too easy and only one mentioned 0-back as too easy. Excluded P43 reported that the 0-back condition got more difficult over the time and that stress from the n-back task was still present during 0-back.

16 of the participants found that the difficulty jump from 2-back to 3-back was higher compared to the jump from 1-back to 2-back. While six said, it is the other way round. P10 even stated that 2-back and 3-back were very similar regarding difficulty.

Approach. Unfortunately, the interview failed to get an idea of how most of the participants approached the task. However, some made interesting remarks. P1 had a technique for 3-back where three letters were memorized then the next three were matched against the three stored letters. This followed a three letter sequence where the participant only guessed and stored the new letters. P19 stated that for the 2-back condition the letters were visualised in contrast to the 3-back where the three letters were said out loud in the head in a rhythm. For P5 stated that a different way of thinking is required for the 2-back and 3-back conditions.

The interview is in line with the TLX and performance results in the sense that all three difficulty levels were different. For the 3-back it was revealed that many had to guess during that condition; however, from the TLX ratings and performance measures it is hard to argue that 3-back was too hard.

After having presented the results of the questionnaires, the performance and interviews it became apparent that the three different levels of difficulty worked as intended. Before the physiological data is used for machine learning an analysis using descriptive and inferential statistics is supposed to reveal which of the three physiological measures were sensitive to the task, thus, make the most sense to be used as machine learning input.

4.2.5 First Glimpse at the Physiological Data

As stated earlier, some participants were excluded and then replaced due to different reasons. In the following, the procedure mentioned at the beginning of how the data was checked is explained along with the reasons making replacements necessary. Finally, the first observations of patterns in the data are discussed.

Data Quality Inspection. After a participant took part in the study, the physiological data was inspected to make sure it can be used for further analysis. Before the inspection, the log files of the different data streams were merged based on the timestamps created with the central data collector. During this process some data was already removed. This includes the first n stimuli per trial where no response could be given, data of the whole practice phase and the data where instructions or similar were presented. The data was labelled according to the difficulty level.

Appendix table A7 summarizes the inspection procedure to get an idea of the quality. The electrodermal activity signal along with the blood volume pulse signal are not included as both were decent for most participants. The pupil data was inspected in the following way. The raw signal was inspected along with different approaches to remove the noise. The quality was categorized to either be a potential candidate to remove (x), very noisy (-), acceptable (o) or good (+). Some signals had a high variance after smoothing indicated by a V. First, the raw signal of the pupil radius of both eyes was inspected. Only for P19, the raw signal was considered to be acceptable for the rest much noise was present. There are different reasons why a sample of the eye tracker does contain no or an unlikely value. If a blink occurs, the pupil cannot be detected as the eye is closed. Further, the detection of the pupil can fail if the eye is half closed or if the viewing angle makes it hard to see the pupil in the camera image. Also, it is possible that the eye tracker detects something else than the pupil, for instance, if mascara is worn. Lastly, the eye tracker cable itself was causing noise in the signal. Hence, as a second step, it was investigated to which extent the signal can be smoothed so that the noise is removed as good as possible while keeping the information of the signal. Each eye was therefore smoothed with a rolling window taking 240 samples using the median that is more robust against outliers in contrast to the mean. In most cases, this improved the signal. As a third step, the signal was inspected by removing outliers and unlikely values, smoothing and averaging the data of both eyes. The eye tracker's SDK provides the pupil size as pupil radius in millimetre and a confidence value. It is -42 if the pupil is not recognized otherwise it is between zero and ten where the latter can be seen as the highest confidence value that the pupil is detected correctly. This data is provided for both eyes. Therefore, all steps were performed on each eye separately. The size of the pupil ranges between 1mm and 8mm [49]

(BNID: 105349), hence, values outside of this range were considered as outliers and removed from the data. Further, samples with a confidence value below five were dropped. Then two smoothing approaches were applied, first the same previously described and second an alternative way by using a Hanning window to convolve the signal. The average of both eyes was computed afterwards and the signal inspected. For the majority, smoothing improved the signal so it could be used.

Excluded Participants, Disruptive Factors. Some participants have been removed from the pupil analysis completely and for some specific trials were removed. P7, P9, P11 and P14 were excluded from the analysis due to the noisy signals. While the smoothing removed much noise, too much information might have been lost in that process. P9 wore mascara even though she was instructed not to. The E4 wristband did not work as well, so no data was gathered with it. For P14 the right eye was very noisy, and the signal was looking okay if only the left was used. For P11 it was the other way round. P7's signal had partly big gaps in the data. All of the participants were replaced with participants using the same counterbalanced order (P7 → P31, P9 → P39, P11 → P35, P14 → 33). P24 was excluded due to an update that interrupted the experiment instead the data of the pilot study participant (P48) was used. P43 was intended to be used as a replacement but was excluded. Noise was not a problem, but it seemed as if P43 was not able to deal with the task very well as indicated by the unusual performance score for which no explanation could be found. Another potential candidate to remove was P3 having similar gaps after smoothing as P7. Most of them were only present in the 3-back condition. Further, P20 who's smart-phone was ringing during the 3-back condition was kept. P2 managed to pull the plug of the monitor with his feet during the first trial of a 0-back sequence. As the effect was expected to be neglect-able, the participant was kept.

Observations during the inspection. Without any statistical analysis, some aspects of the pupil data could already be observed. For all conditions the difference between 0-back and the subsequent n-back was visible. Further, for most a difference between the 1-back condition and the two other conditions was visible. Similar to the pretest this difference was not that visible when comparing 2-back and 3-back. Another aspect that was noticed is that for the 1-back task often trial #1 was different compared to the remaining three trials leaving room for speculation. As they were told to give their best and stay concentrated in the first trial they might have put more effort into performing well while in the remaining trials they might have learned that not that much effort is necessary for the 1-back task.

During the inspection, it was hoped that the two other physiological signals coming from the E4 might show some interesting pattern that could be investigated further. Unfortunately, that was not the case. The skin conductance level (SCL) increased, for some participants, over time independent of difficulty level. Most likely this is caused by sweating more and an increased temperature where the electrodes were placed. For some participants, at the beginning of a difficulty level, the SCL was high and gradually going down until the end of the condition. For some, the SCL appears to be very similar in every condition. In contrast to the pupil, no repeating pattern resulting from the modulation of the difficulty could be observed. The heart rate estimates as well did not show a clear and repeating pattern.

4.2.6 Pupil Measures

Before even starting to analyse the signal using inferential or descriptive statistics the inspection of each participants data already indicated the differences between conditions, most notably between 1-back and 2-back. Further, during the inspection, it became clear that the data was often very noisy and smoothing is required. Appendix B contains the tables with the complete analysis' results.

Preprocessing. Smoothing was applied which removes outliers and accounts for missing values. The rolling window function of the pandas library was used with a window size of 240 samples and a minimum window size of 1 sample. The median of these windows was then used as smoothed data points. In general, the pupil data was quite noisy; hence, two times the sampling rate of the eye tracker was used for the windows which roughly corresponds to two seconds. The median was chosen due to the many outliers and missing values. Even though the smoothing was applied, at some data points there was still a minimal amount of missing values which were then replaced with backward filling. It needs to be noted that by smoothing the information of blinks is lost.

If there is data for both eyes, there are two possibilities. Either the average of both or the eye with better data can be used. Except for five participants, the average of both eyes was used as final pupil data (as done in [82]), for the others, the signal was much better when choosing the better eye. Better, in this case, was decided by inspecting the raw data of each eye. The average was chosen, as deciding which eye is better requires to define some metric. Every participant's data was inspected to check the quality manually resulting in using both eyes for almost all participants. Still, a metric was used. As a step to decide for the better eye the following steps were taken. First, it was tested if the expected pattern was visible using the mean of each difficulty level. It was checked if the mean of 1-back is smaller than the mean of the 2-Back condition and the latter was compared to the 3-Back condition. In the best case 1-Back < 2-Back < 3-Back was true. For some only 1-Back < 2-Back was true, for some only 1-Back < 3-Back and for some both held. Then it was as well checked whether the difference between conditions was smaller than 0.08 and 0.1. In cases where the results (see appendix table A8) were bad taking only one eye improved the visibility of the pattern. P3, P10, P11, P14, P20 profited from this procedure - P11 and P14 have been excluded from the analysis as described earlier.

Differences Between Conditions. The analysis investigated whether several statistical features of the pupil radius of the three difficulty conditions were significantly different. The median (MED), interquartile range (IQR), quantiles 0.25 and 0.75 (Q25, Q75), mean (M), standard deviation (SD), maximum (MAX), minimum (MIN) and the peak to peak value (P2P) inspired by [82] and [64] were computed for each difficulty level. The statistical features were calculated on the smoothed signal but also on a calibrated version. The mean of the 0-back condition was used as a baseline measure, similar to [82], to calibrate each trial run using the formula: $\frac{x_{pupilRadius} - mean_{n0}}{mean_{n0}}$. Zhou et al. [82] used a resting phase as a baseline to calibrate the data of the task that followed the resting phase. They were dealing with electrodermal activity data. Still, physiological data in general is often investigated by comparing baseline measures with those of the experiment's condition [24], for example, see Wilson and Russell [78].

The results will be summarized for all the features, all descriptive statistics along with the ANOVA results and the corresponding post hoc tests can be found in appendix B.

All Friedman tests and ANOVAs were significant for the three conditions except for IQR $\chi^2_{IQR} = 4.333, p < 0.115$ when using 0-back calibration. All post hoc tests were significant when comparing 1-back and 2-back except for SD $T_{SD_0} = 2.339, p = 0.028$ when using 0-back calibration. All post hoc tests were significant when comparing 1-back and 3-back except for P2P $T_{P2P_0} = 2.121, p = 0.045$ when using 0-back calibration. Except for MAX $T_{MAX} = -2.717, p = 0.012$ all post hoc tests were not significant when comparing 2-back and 3-back. As an example the mean is reported in more detail. For the mean of the pupil radius ($M_{1back} = 3.143, SD_{1back} = 0.324$ $M_{2back} = 3.380, SD_{2back} = 0.271$ $M_{3back} = 3.415, SD_{3back} = 0.295$) the ANOVA $F(2, 23) = 85.710, p < 0.001, \eta^2_p = 0.788$ revealed that there are significant differences between the difficulty conditions. Paired t-tests were used as post hoc tests with an adjusted significance level $\alpha = 0.05/3 = 0.0166$ showing significant results when 1-back is compared with the other two (see table 7).

			No Calibration			0-back Calibration		
			T	p	Cohen's d	T	p	Cohen's d
1-back	-	2-back	-11.901	< .001	-2.429	-8.706	< .001	-1.777
1-back	-	3-back	-10.107	< .001	-2.063	-9.938	< .001	-2.028
2-back	-	3-back	-1.717	0.099	-0.35	-0.301	0.766	-0.062

Table 7. Pairwise comparison with the paired samples t-test for the mean pupil radius.

When using the 0-back calibration approach similar significant results were found. The means of the difficulty levels were $M_{1back} = -0.007, SD_{1back} = 0.023$, $M_{2back} = 0.060, SD_{2back} = 0.040$, $M_{3back} = 0.063, SD_{3back} = 0.036$. The repeated measures ANOVA resulted in $F(2, 23) = 54.669, p < 0.001, \eta_p^2 = 0.704$. The results of the post hoc tests were significant when 1-back is compared with the other two as seen in table 7.

As an interim result, it can be stated that the pupil was sensitive to the difficulty level. While the most statistical features of the radius were not significantly different between 2-back and 3-back, in both approaches with and without a calibration procedure, the differences between 1-back and 2-back or 3-back were significant.

4.2.7 Cardiovascular Measures

Traditional Measures. Previously the cardiovascular activities were introduced. Before continuing with the analysis of the results some remarks have to be made regarding the measures. The E4 provides a blood volume pulse (BVP) signal in nanowatt (nw), an estimate of the heart rate (bpm) and the reciprocal value called interbeat interval (IBI) in seconds. The latter is the time which has passed between two adjacent heartbeats. As a reminder, heart rate variability (HRV) describes the change between two adjacent heartbeats, and a decrease in HRV is related to an increase in cognitive demand. A decrease in HRV is related to an increased HR. It has to be kept in mind that HRV is usually derived from an ECG signal, however, in this case, IBI is derived from the BVP signal. Moreover, one cannot say that deriving from the BVP signal yields the same results as if derived from the ECG signal. For short-term HRV component analysis usually at least two minutes of data are required [43]. There are about 60 seconds of data per trial run, leading to four minutes per difficulty level. Malik [43] discusses the different measures of HRV, regarding time-domain analysis two measures are interesting. First, the standard deviation of all NN intervals (SDNN) - NN, in this case, refers to the normal-to-normal-interval of the QSR component of the ECG signal. The IBI will be used as a substitute of an NN interval.

Second, the square root of the mean differences of successive NN intervals (RMSSD). Again we can use the IBI as a replacement for the NN intervals. The measures under investigation therefore were: heart rate (HR), interbeat interval (IBI), their variability (HRV, IBIV), the variability's standard deviation (HRVSD, IBIVSD) and the square root of the mean differences of successive IBI intervals (RMSIBI). Note that the variability describes differences of successive intervals.

The measures were calculated for each difficulty level leading to 24 samples for each. P24's data was used to account for the missing data of P48 who was used for the pupil analysis. Appendix D contains the tables with the complete analysis' results.

There were no significant results for most of the measures. Only HR ($M_{1back} = 74.737, SD_{1back} = 9.721$ $M_{2back} = 76.431, SD_{2back} = 10.997$ $M_{3back} = 76.166, SD_{3back} = 10.09$) and IBI ($M_{1back} = 0.820, SD_{1back} = 0.110$ $M_{2back} = 0.804, SD_{2back} = 0.113$ $M_{3back} = 0.806, SD_{3back} = 0.108$) are worth mentioning. The ANOVA results $F_{HR}(2, 23) = 2.493, p = 0.094, \eta_p^2 = 0.098$ and $F_{IBI}(2, 23) = 2.744, p = 0.075, \eta_p^2 = 0.107$ did not show any significant effects.

The results are not too surprising when keeping in mind that the effects on traditional heart-related measures usually can be seen much better over longer periods. It can be speculated that if longer periods per participant are recorded a difference might become significant. However, it can be said that the n-back task with its different difficulty conditions did not show a significant effect on any of the measures.

Peak BVP Measures. The work by Zhou et al. [80] was mentioned previously as they investigated whether the raw BVP signal can be utilized for short time periods. They propose measures regarding the peaks of the signal which will be used for analysis of the raw signal as well. For preprocessing, following their approach, a Hanning window with a size of 200 was used to convolve the signal. Then the z-score was used for normalization. The smoothing and normalization were applied to the complete data of a participant. The results did not show any significant results, for instance, the peak count ($M_{1back} = 143.875, SD_{1back} = 43.249$ $M_{2back} = 136.458, SD_{2back} = 35.743$ $M_{3back} = 139.375, SD_{3back} = 35.038$) had a difference between the three conditions which was not significant. For the peak detection *scipy.signal.find_peaks* was used. Between two peaks there had to be at least 32 samples (half of the sampling rate of the BVP signal).

Neither the traditional measures nor the experimental peak measures did show any significant differences among the three difficulty levels.

4.2.8 Electrodermal Activity

For the electrodermal activity (EDA) there are the two main measure of skin conductance level (SCL) and the skin conductance responses SCR as seen in previously in figure 3. To investigate if the EDA signal was sensitive to experimental evaluation the following measures were analysed using repeated measures ANOVA: the mean SCL (mSCL), the number of SCR's determined using peak detection (SCR), the maximum of the peak SCL values (maxSCR), mean of the peak values and their variance (mSCR, varSCR). P24's data was used to account for the missing data of P48 who was used for the pupil analysis. Appendix C contains the tables with the complete analysis' results.

The first preprocessing step was to apply a Bartlett window to remove the trend of increasing SCL over time [26]. Then due to the subject-to-subject differences, the data was standardized. As an alternative, to account for the differences, the 0-back condition was used as described in the pupil analysis. The steps were performed on all data of a single participant.

	Mean SCL			Number of SCR		
	1-back	2-back	3-back	1-back	2-back	3-back
Mean	-0.331	0.240	-0.004	23.917	27.333	24.250
Std. Error of Mean	0.219	0.193	0.156	1.434	1.221	1.624
Median	-0.272	0.195	0.028	24.500	29.000	24.500
Std. Deviation	1.071	0.947	0.765	7.027	5.983	7.958
Range	5.233	4.332	3.113	26.000	28.000	28.000
Minimum	-3.606	-1.349	-1.561	10.000	10.000	10.000
Maximum	1.627	2.983	1.552	36.000	38.000	38.000

Table 8. Descriptive statistics of two of the EDA measures: mean SCL and number of SCRs. These are the values for the variant were the data was normalized using the z-score.

In table 8 the descriptive statistics of the mean SCL and the number of SCRs are presented. For the SCRs the ANOVA ($F_{SCR}(2, 23) = 3.805, p = 0.03, \eta_p^2 = 0.142$) indicated that there were significant differences; however, the post hoc tests did not yield any significant results. Analysis of the other measures resulted in no significant differences.

Similar to the cardiovascular measures the EDA measures did not show significant results indicating that the three different n-back levels did not elicit patterns in the physiological responses that might be useful.

4.2.9 Discussion & Summary

The task difficulty levels were perceived as expected indicated by the performance measures and the NASA-TLX scores. 1-back was less demanding than 2-back and 2-back less demanding than 3-back. The interviews revealed that 3-back challenged the participants more as they had to guess when losing track of the sequence. Looking back at related work where n-back was used as well not much was reported whether 3-back was too hard or if participants were guessing. However, in the present n-back variant feedback was given that affected the participants. When knowing that they made a mistake, they were able to adjust. While feedback might have been beneficial in some situations (e.g. when they forgot the letter presented during the 0-back instruction), it might have caused additional load or stress.

For the EDA and heart-related measures, no significant effects of task difficulty could be found. As stated earlier analysing heart-related measures is more suitable for a longer time span, thus, might be more sensitive in longer periods than one four-minute block of data of each difficulty level. Comparing the BVP results with Zhou et al.'s [80] approach, it can be said that they used a dual-task and measured the BVP signal at the middle finger in contrast to the E4 used in the present study measuring it at the wrist.

Concerning the EDA result, only speculations can be made why it was not sensitive to the task. The number of specific SCRs might not have been significantly different due to the short period of the task. Also, the non-specific SCRs are included in the number. With the smoothing it was attempted to account for an increasing SCL level; however, using a longer baseline period before each task might have been more fruitful.

In contrast to the EDA and HR/BVP measures the pupil was sensitive to the task difficulty, especially visible for 1-back compared with the other two. In contrast to the performance and NASA-TLX, however, significant differences between 2-back and 3-back were not found except for the maximum of the pupil radius per difficulty level. Both conditions resulted in insignificantly different pupil responses. It can be speculated that both produced a highly similar cognitive load level which was measured objectively with the pupil. However, according to the subjective data they were perceived as different levels (NASA-TLX, interviews). The 3-back performance does, however, not support this theory if it is interpreted as a measure of cognitive demand.

The n-back task was used to produce different levels of cognitive state which was done successfully. It remains an open question why some of the physiological measures did not respond to the difficulty levels. The pupil measures results are promising, and it is expected that the data, when used for machine learning, can discriminate at least between the 1-back condition and the other two.

4.3 Summary

The data gathering study collected pupil data of 24 participants which now can be used as input for machine learning. The n-back task difficulty levels were perceived as three different levels one more difficult than the other. However, the difference was not visible between 2-back and 3-back in the pupil data. The other measures did not show any interesting results.

With the collected pupil data it is now possible to build predictive models across all users which is subject of the next section.

5 CLASSIFYING COGNITIVE STATE ACROSS USERS

With the analysis of the data gathering study, the next step is to use the pupil data as input for machine learning. By doing so, the first research question is addressed which is concerned with using physiological data to distinguish between the different difficulty levels of the n-back task using predictive models trained with data of all subjects. One outcome of the analysis was that only pupil data of 1-back and 2-back were significantly different. Hence, the classification performance is expected to be not as good when 2- and 3-back are included. Further, the other two measures are left out as they did not seem to be sensitive to the difficulty levels.

In [65] a typical machine learning workflow, to create a model that can be used for prediction, consists of the three phases of preprocessing, learning and evaluation. The preprocessing usually includes steps such as the labelling of the data, feature extraction, scaling and selection. In the learning phase model selection and cross-validation is done. In this phase, the best approach to process the data along with the most suited learning algorithm is attempted to be found. The last step of evaluation usually requires to take out a part of the training data in the very beginning to test the final model with it.

This section focuses on the two first steps; hence, the processing of the pupil data will be described along with the description of the model selection process. To evaluate the final model no participant data was excluded in the very beginning. Instead, the data that was gathered in the second study was used to evaluate the final model's performance which is discussed in section 6.3.

This section starts with a description of the supervised machine learning algorithm of choice - the Random Forest Classifier (RFC). The goals and approaches of using the RFC for classification are discussed. Then the preprocessing, and the description of the learning phase are presented. Results are presented subsequently and discussed.

Data preprocessing, analysis and machine learning was done with *Python* using *NumPy* [1], *pandas* [2], *SciPy* [3] and *scikit-learn* [4] [61].

5.1 Random Forest Classifier

Related work did not reveal an algorithm that performs best when utilizing physiological data or specifically pupil dilation. As a result, the RFC was chosen as it

[1] http://www.numpy.org/ (Accessed: March 4, 2018)

[2] https://pandas.pydata.org/ (Accessed: January 21, 2018)

[3] https://www.scipy.org/ (Accessed: January 13, 2018)

[4] http://scikit-learn.org/stable/ (Accessed: January 13, 2018)

works very well for "general-purpose classification" [11] and has the advantage of being robust against overfitting [14]. Additionally, it performed well in [64] where pupil dilation was used as input.

Before describing the approach to use the classifier for the n-back task, a short introduction is given based on [14]: The Random Forest Classifier (RFC) is a so-called ensemble technique which trains multiple decision trees and makes a classification based on the majority vote of all trees' predictions. In scikit's implementation, each tree in the forest has a probability value for each class. Thus, the average output value of all trees is used to decide the final output instead of using voting.

A single decision tree in the forest starts with the root node containing all samples. Then a set of features is randomly selected that are used to decide how the current node's samples can be optimally split into two new nodes. For instance, the features might be the mean and median pupil radius. The tree will now first use the mean an try to split the node's samples into two sets (two new nodes) by putting all samples with a mean radius > 3.1 into the first and the rest into the second. For each new set, a Gini impurity value can be computed. Impurity describes how impure a set is in the sense of how many different labels (e.g. three cognitive states) does the set contain and how high is the chance of incorrectly assigning a label to a sample of the node's set. If all samples have the same label, the impurity is zero. If there are more than two labels, the impurity depends on the number of samples per label. The tree algorithm will use the impurity value to decide how the samples at a node can be split and it will compare the impurity using the mean pupil radius versus using the median and potentially other features. The process is repeated until a minimum number of samples is reached which then are the leaves of the tree.

From the short explanation of a decision tree and how it conceptually works, several relevant parameters for the RFC can be derived. Among them are the number of trees, the number of random features used when splitting a tree, the number of samples in a leaf, the tree's depth and whether to use bootstrap samples. Each tree in a forest uses a subset of all available samples. Typically, the samples are drawn with replacement so-called bootstrap aggregating or bagging. The latter comes with the benefit that a subset of the training data is left out which can be used as a validation set. Thus, an RFC computes a generalization error or out-of-bag error (OOB) that can be used for validation. Additionally, the RFC keeps track of how good single features were able to split the samples into sets with lower impurity, hence, giving insight on the feature importances. Decision trees do not need feature scaling and work very well with a large number of features even if they are correlated.

After this short introduction to RFCs and decision trees, the goals of the classification besides the cognitive state assessment, are highlighted.

5.2 Goals & Approaches

The first research question is concerned with classification across users. With the three difficulty levels of the n-back task, there are three classes between which a classifier is supposed to discriminate. Further, the pupil radius will be the only measure that is going to be used as input for the classifier. In the following goals of utilizing machine learning and the way they are approached are described.

Main Goal: cross-user classification for the n-back task using pupil radius.

Goal 1: Specificity. Having three classes the performance of making 3-class (1-back vs 2-back vs 3-back) and 2-class classification (1-back vs 2-back, 1-back vs 3-back, 2-back vs 3-back) is investigated. Classification performance is expected to be worse when 2-back and 3-back are involved as the inferential statistics' analysis did not reveal any significant differences between the two conditions.

Goal 2: Window Size. A trade-off between window size and performance was discussed in the related work section. Thus, it will be investigated whether this as well counts for the pupil dilation and the present approach. The following window and step sizes (in seconds) - inspired by [31] and Rajan et al. [64] - will be compared: (3, 1), (5, 1), (10, 1), (30, 1), (60+, 1). The largest window size of 60 will use the data of a whole trial run, thus, will contain slightly more than 60 seconds. Further, the window size is relevant for future approaches where a real-time system with online classification is built. A larger window size will require more processing time than a smaller window.

Goal 3: Individual Differenecs. Physiology differs among individuals, for instance, for P1 the mean pupil radius during the 1-back task was 3.47 in contrast to P4 with a mean of 2.63. Thus, there is a need to account for these differences. An option is to scale (or normalize) each participant's data as the scaling converts the pupil data of individuals to the same range. This approach is very similar to Zhou et al.'s [82] preprocessing where the z-score was used on smoothed data to "compensate for differences between participants". They applied this processing on GSR and pupil data.

Scaling requires the mean and standard deviation over a given time, thus, computing these values on the complete data of a participant is not possible in a real-time scenario. For that case, a calibration phase is required to compute these values. As a result, an alternative to the scaling per participant is to use the 0-back trial phases as done during the statistical analysis in the last section.

With scaled or calibrated input better performance is expected.

Goal 4: Utilizing pupil radius. As the statistical analysis did not indicate a significant effect of task difficulty on the EDA and BVP measures only the pupil radius is used. Getting all possible information that might be included in the pupil radius signal is, therefore, a goal which is expected to result in better classifier performance. For instance, the inferential analysis comparing 2-back and 3-back only revealed a significant difference of the maximum pupil radius. Thus, it makes sense to use multiple statistical features as each might contain different information potentially increasing performance. The same nine statistical features investigated during inferential analysis are used (mean, median, interquartile range, quantiles 0.25 and 0.75, standard deviation, maximum, minimum and the peak to peak value). Further, Rajan et al. [64] computed features on three different signals of the pupil data: the main signal ($x[n]$), the derivative signal ($x[n+1]-x[n]$) and the percentage change signal ($\frac{x[n+1]-x[n]}{x[n]} * 100$). As they had a good result with the approach, the features were as well computed for the two new signals leading to a total of 27 features.

This allows investigating which features were most important when training the RFC and if utilizing features of the different signals has a positive effect on performance.

Goal 5: Validation. In order to be able to argue that the classification results are generalizable a validation of the results is necessary. As stated earlier the RFC already provides an error estimation when bootstrapping is used - the OOB. However, if future work wants to compare the RFC with another classifier which does not have a built-in error estimation and cannot be trained and tested with the same data sets left out during bootstrap aggregation, the OOB error is not suitable. Therefore, to validate the results cross-validation (CV) is used.

For physiological data that can differ a lot from individual to individual and with the goal to have a model that works for all users the approach to use is, according to [30], leave-one-user-out-cross-validation (LOUOCV). The idea is to train multiple models with different training and validation sets. Thus, in each CV run the data of one subject is taken out, the model is trained with the data of the remaining 23 participants and tested with the left-out-subject. With 24 participants this can be repeated 24 times. In [30] the context in which the approach is discussed is analysing brain architecture. It can be assumed that also other physiological data should be treated the same. For instance, Rajan et al. [64] as well used the LOUOCV for their classifiers using pupil data as input. Leave-10% of users out was used in [70] and common LOUOCV in [81].

As an additional form of validating classifier performance, the performance of individual models can be compared to the population models. This also allows comparing the approach with related work where no population models were built. Individual models, however, can not make use of the same CV approach, hence, instead of leaving out a participant the trial runs are left out, e.g. all first trial runs of each difficulty. However, it has to be kept in mind that the amount of data trained on is quite small when using only the data of an individual. There are roughly four minutes per difficulty level).

Performance of a classifier was mentioned multiple times; hence, it is necessary to decide on a specific metric that is used to be able to compare different approaches. The accuracy describes how many of the given training set samples are classified correctly. Assuming there is training data for cognitive state A and cognitive state B. The accuracy would be the percentage of correctly classified states A and B. There is also the precision and recall metric where precision describes the ability not falsely to classify a B sample as A sample. While recall describes the ability to classify all A samples as A samples, independent of how many mistakes are made (B samples classified as A samples). The f1-score describes the weighted average of precision and recall. Which metric to use for evaluating the results depends on the given problem. As an example imagine the classification result would trigger a car to take over the breaks. With a high recall ability, no situation would be missed that truly requires the car to use the breaks automatically. With a high precision; however, some of these situations might also be missed those that will lead to an accident. A user interface which is adapted to the user state, having a high recall might lead to adapting too much and too often. All metrics were computed while training the models. To compare the models only accuracy was used as the use case of classifying the n-back task difficulty levels is very abstract and deciding which metric makes the most sense is only possible if the action triggered by the classification result becomes concrete.

After discussing the goals and approaches the preprocessing of the data is described next.

5.3 Preprocessing

In the preprocessing phase (see figure 8) the pupil data of each participant was labelled by using the different log files created by the central data collector. The log file with the information about the state of the experiment (e.g. which trial run, which conditions and more) was merged with the pupil data based on timestamps. After this merge, the first unnecessary data was removed: data of the practice phases and the data which relates to the time where letters were already presented but no answer could be given (e.g. during the first n stimuli).

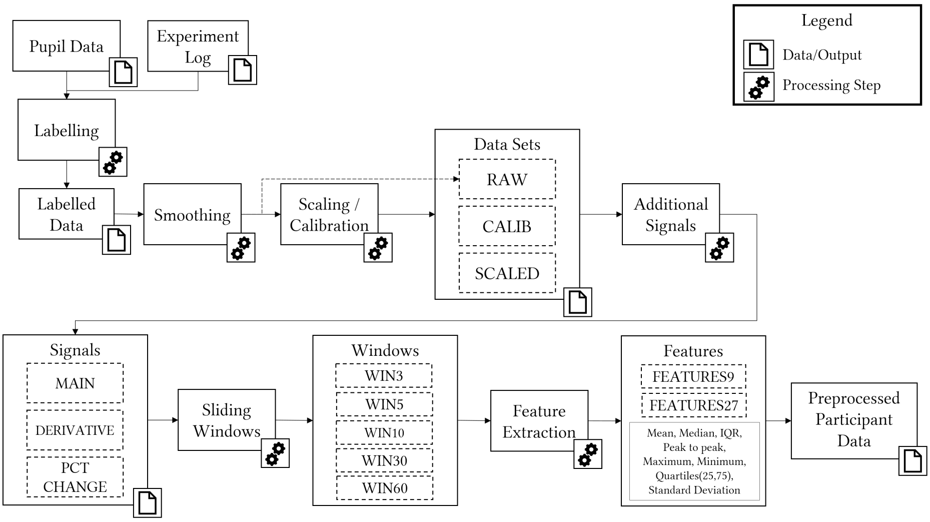


Fig. 8. The preprocessing pipeline done for each participant separately.

If the eye tracker does not recognize the pupil or the eye is closed because of blinking, samples will have an undefined pupil size - a missing value per sample. It is assumed that the sampling rate of the eye-tracker is constant, this is exploited to create sliding windows that are discussed later. As a result, it was no option to drop samples with missing pupil data. An option would have been to interpolate the values or fill them with a forward or backward fill method such as *pandas.DataFrame.fillna()*. However, as a second issue had to be solved, namely, removing outliers, smoothing was used to address both, the missing values but also outliers without removing any samples so it could be assumed that 120 samples corresponded to one second of data send by the eye-tracker with a rate of 120 Hz. As depicted in figure 5 in the last section one trial run had a 0-back phase and an n-back phase. Each of these two phases of the four trials of a trial run was smoothed separately. The smoothing was the same as applied during the analysis presented in the last section.

After smoothing a separate data set was created using the 0-back calibration approach followed by creating another data set which contained scaled data (see Goal 3). Scikit's *StandardScaler* was used that computes the z-score which removes the mean and scales the data to unit variance. The smoothing already accounted for outliers. Hence, the *StandardScaler* should be fine to use, and it is not necessary to use another scaling method which is more robust against outliers. As a result, three different data sets were created: 0-calibrated, scaled and raw. The three sets will be referenced to as *CALIB*, *SCALED* and *RAW*.

For each of these sets, features were extracted using sliding windows which is a common way when dealing with physiological time series data as it enables to capture temporal changes [64]. As stated earlier (see Goal 2) features were extracted from different windows. To compute the different windows 120 samples (the sampling rate of the eye-tracker) were considered as one second, hence, for a ten-second window $10 * 120 samples$ were used. As mentioned earlier, it is assumed that the sampling rate was constant which makes this approach reasonable. Now it is also evident why outliers and missing values could not be dropped as this approach requires 120 samples to represent one second. The windows were created for each trial run separately so that no jumps from one trial run to the next influenced the windowing. Incomplete windows were ignored. The windows will be referenced to as, e.g. $WIN5$ for windows with a length of five seconds. The statistical features mentioned in Goal 4 were extracted on the three different signals ($MAIN$, $DERIVATIVE$,$PCTCHANGE$). These will be referenced to as $FEATURES9$ if only the statistical features of the main signal are used and $FEATURES27$ if all are used. With the preprocessing per participant data sets were created that can either be used for building individual models or population models. The next section is concerned with the learning phase and gives an overview of the approaches.

5.4 Learning

The cross-validation approach for both individual and population models was described earlier (see Goal 5). Using only the RFC as algorithm the choice of parameters is briefly discussed.

Breiman [14] indicates that using bootstrap aggregation results in better performance on the contrary Louppe [41] states that it is not crucial to "obtain good accuracy" and that not using bootstrap samples results in better performance. As the LOUOCV is used the OOB error estimation is not required, and thus, bootstrapping can be disabled. Scikit's user guide gives suggestions on the number of features randomly selected when splitting a node, namely, for classification $\sqrt{\text{number\_of\_all\_feauters}}$. When using all 27 features described in Goal 4, nine features are selected randomly in every attempt to split a tree. Boulesteix et al. [12] explain that considering too many features will most likely lead to choosing only the features containing the most information and leaving out moderate effect features that still can add to the information gain. In contrast, considering a small amount per split might lead to never selecting the features that contain the most information. The RFCs trained were using the suggested value by the user guide. As stated by Breiman [14] the number of trees will affect the generalization error as it will decrease with more trees. Scikit's user guide states that "The larger [the number of trees] the better [...] results will stop getting significantly better beyond a critical number of trees". In a later stage the parameter might be tuned, for the first five goals, however, the default value of scikit's RFC is used (10 trees) with a second RFC using 50 trees. Further, the default values for maximum depth (that is none) of a tree and the number of samples required to split a node are used (that is

2) as suggested by the user guide.

Two different versions of the RFC using different amounts of trees (*RFC*10, *RFC*50), three different preprocessed data sets (*CALIB*, *SCALED*, *RAW*), five window sizes (*WIN*3, *WIN*5, *WIN*10, *WIN*30, *WIN*60), two feature sets (*FEATURES*9, *FEATURES*27), three 2-class (1*v*2, 1*v*3, 2*v*3) and one 3-class classifications (1*v*2*v*3) are used. Which results in training 240 different pipelines each validated using CV. This allows investigating the specificity of classifying between the three classes (Goal 1), the effect of window size (Goal 2), if individuals differences can be accounted for (Goal 3) and the effect of pupil features (Goal 4). Afterwards, the best approaches might be taken and tuned further.

5.5 Results

The primary goal was formulated as the successful discrimination of difficulty levels independent of the individual - cross-user classification. Between which levels, it is possible to discriminate well is investigated by looking at the specificity (Goal 1). In the following, the abbreviation *SC* is used describing either 2-class classification or 3-class classification. With Goals 2 to 4 (window sizes, individual differences, utilizing pupil) it is possible to investigate what affects the performance of each *SC* classification. Those are going to be referred to as *FACTORS* . For the *FACTOR* window size the reference *WIN* is used, *INDIV* for the individual differences (Goal 2) and *FEATURES* for the different sets of features used. As a first attempt towards tuning the RFC was trained using ten and 50 trees, this is another *FACTOR* referred to as *RFC* .

As stated in the previous section 240 pipelines are a result of combining *SC* , *RFC* , *WIN* , *INDIV* and *FEATURES* . First, the specificity is investigated by grouping the 240 pipelines by *SC* leading to 60 pipelines for each. Second, the grouping is kept, and the influences of each *FACTOR* on the performance of *SC* is analysed using inferential statistics. For simplicity, when analysing a single *FACTOR* , it is assumed that it is not affected by the others. For instance, it is assumed that the effect of the choice of features is independent of the choice of window size or the number of trees. This has to be kept in mind when interpreting the results. Third, the population model results are contrasted with individual models. Lastly, a look is taken at pipelines with good performance to see what accuracies are possible.

The inferential analysis makes use of non-parametric tests: Friedman test if the *FACTOR* has more than two levels with a significance level of 0.05, followed by pairwise Wilcoxon tests as post hoc analysis with an adjusted alpha of $0.05/factorLevels$. Otherwise, the Wilcoxon test is used without an adjustment.

Non-parametric tests do not assume that the samples are independent which each pipeline score is not due to being trained with the same data. Further, for the Wilcoxon test paired samples are required which is given by pairing the performance scores of the combination of $FACTORS$. For instance, when investigating WIN a pair would be pipeline ($RFC10$, $FEATURES27$, RAW, $1v2$, $WIN5$) and ($RFC10$, $FEATURES27$, RAW, $1v2$, $WIN60$).

Only parts of the analysis' results are presented in this section for a more detailed view see appendix E.

5.5.1 Goal 1: Specificity

Figure 9 shows the mean classification accuracy for each SC independent of $FACTORS$ the descriptive statistics can be found in table 9.

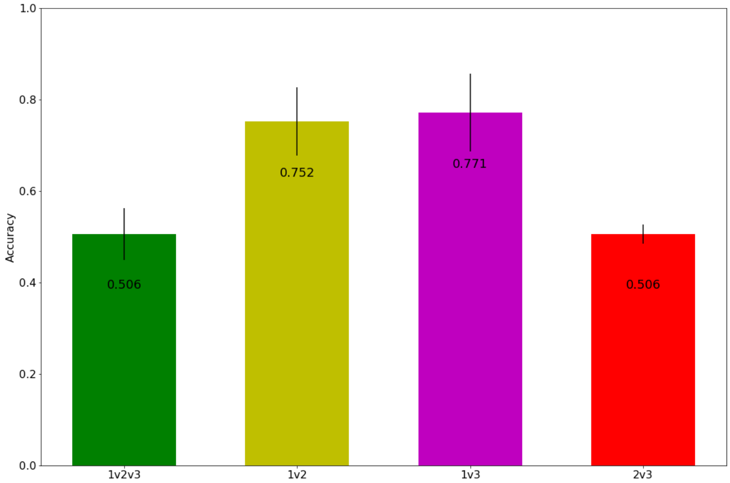


Fig. 9. Mean classification accuracy for each SC independent of $FACTORS$

Classes	Count	Mean	SD	Min	Max	Q25	Q50 (Median)	Q75
1v2v3	60	0.506	0.057	0.411	0.62	0.457	0.503	0.55
1v2	60	0.752	0.074	0.618	0.891	0.685	0.75	0.814
1v3	60	0.771	0.085	0.613	0.901	0.678	0.795	0.839
2v3	60	0.506	0.021	0.456	0.583	0.494	0.505	0.512

Table 9. Descriptive Statistics Specificity

It can be seen that the performance when 2-back and 3-back are part of the classification ($2v3$, $1v2v3$) is around 50% on average which is not better than a coin flip. However, for $1v2v3$ one pipeline reached an accuracy of 0.62 indicating that better results might be possible. The analysis of the data gathering study did not reveal significant differences between the two higher difficulty levels, except for the maximum. It seems that the RFC was not able to learn the subtle differences that might exist between these conditions.

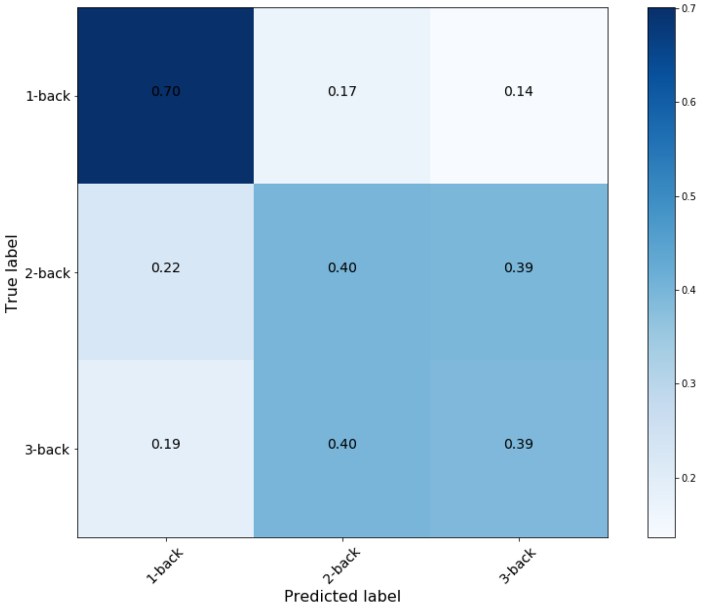


Fig. 10. Normalized confusion matrix for $1v2v3$

The confusion matrix of $1v2v3$ is depicted in figure 10. 1-back was classified correctly in 70% of the times. However, the discrimination between the other two classes does not appear to be possible with the given pupil data. Discrimination between 1-back and the more difficult levels seems to work reasonably well as the descriptive table shows. One pipeline reached 90% accuracy for $1v3$.

With the average over all pipelines per SC, it can be seen that at least 2-class classification involving 1-back performs well across users. Next, it is investigated what affects the performance positively.

5.5.2 Goal 2: Window Size

The mean accuracies for each window size and *SC* is depicted in figure 11, the appendix includes all statistics (see E.3.1).

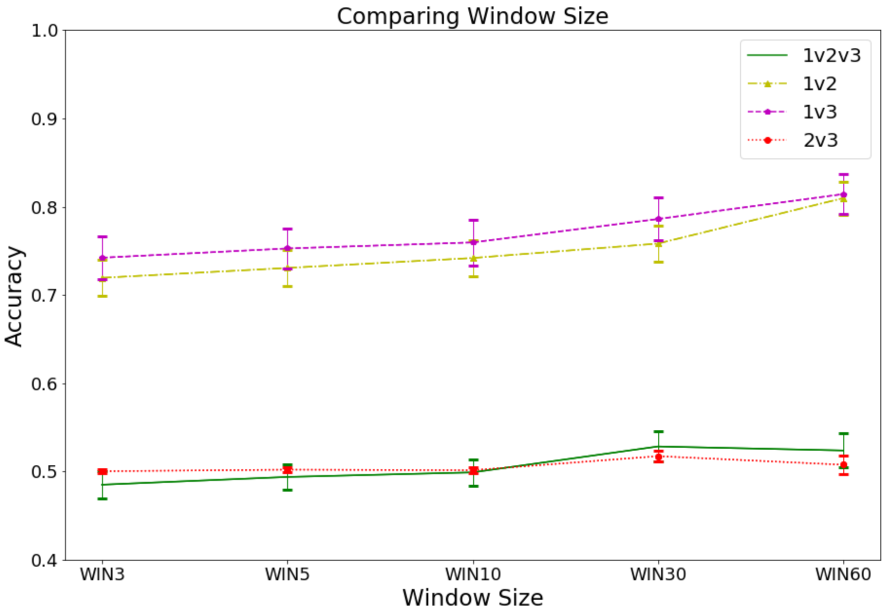


Fig. 11. Mean accuracies for each window size with standard deviation bars.

The trend of increasing accuracy with increasing window size can be observed, however, for the bad performing $2v3$ and $1v2v3$ a decrease seems to happen from $WIN30$ to $WIN60$. It has to be noted that the 60-second window approaches have only one sample per trial run available ($4 * 24$ samples per class) in contrast to $WIN30$ with more than 3000 samples per class due to the sliding window approach. Even with small sizes over 70% accuracy was reached indicating that even these small time frames can capture state changes.

Each *SC* was analysed using a Friedman test indicating significant differences between all *WIN* per *SC*. Post hoc analysis for $2v3$ did not reveal any significant differences in contrast to $1v2$ and $1v3$ where all pairwise comparisons were significant except for the $WIN10, 1v3$ and $WIN5, 1v3$. For $1v2v3$ comparing $WIN60$ with the others only was significant compared to $WIN3$. It has to be noted that the results of the Wilcoxon test have to be taken with caution as per *WIN* only 12 samples are available for each *SC*. The scipy version of the test (*scipy.stats.wilcoxon*) uses the normal distribution for calculating the p-value which usually requires a larger number of samples (e.g. more than 20). If this fact is ignored the results might indicate that the drop of performance from $WIN30$ to $WIN60$ is a random occurrence and might be caused by the general bad performance of the 3-back involving classifications.

With the average over all pipelines per SC and WIN and the inferential analysis comparing the differences it can be assumed that increasing window size leads to an increase in performance. Next, it is investigated how accounting for individual differences can be done.

5.5.3 Goal 3: Individual Differences

Two different approaches for accounting for differences were used for training - $CALIB$ using the 0-back condition to scale the data and $SCALED$ which uses all available data of a participant for scaling. The third data set created in the preprocessing phase was RAW which does not account for the differences at all. Thus, the latter is expected to yield worse performance results than the others but will give a lower bound for the classification performance.
The mean accuracies for each $INDIV$ are depicted in figure 12, the appendix includes all statistics (see E.3.2).

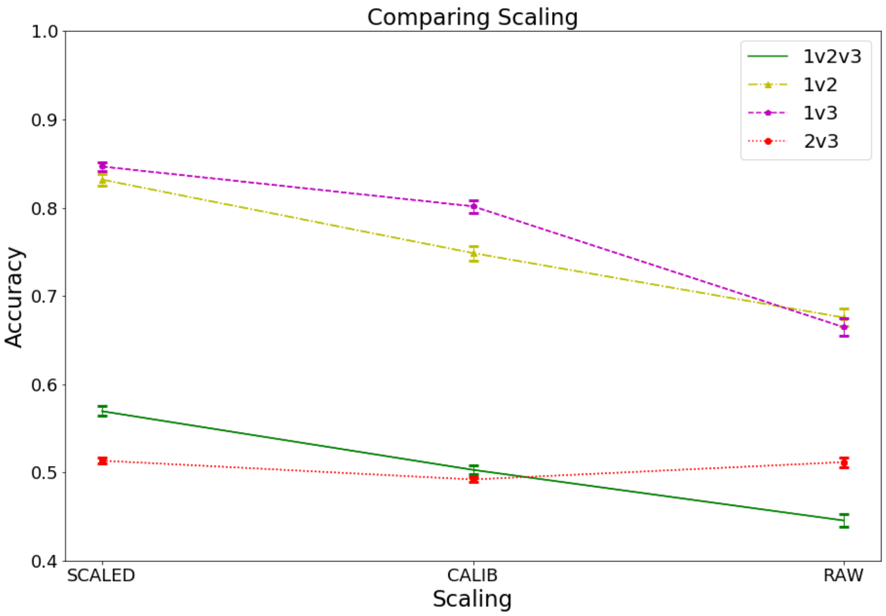


Fig. 12. Mean accuracies for each $INDIV$ with standard deviation bars.

The general tendency that can be observed is that *SCALED* performs better than *CALIB* which performs better than *RAW*. The exception is $2v3$ where the latter performs as good as *SCALED* and *CALIB* is the worst performer.

Friedman tests were significant for each *SC* comparing the three *INDIV* . Post hoc tests were significant except when comparing $2v3$, *RAW* and $2v3$, *SCALED*. The number of samples compared where 20 per *INDIV* .

Between the mean of *SCALED*, $1v3$ and *RAW*, $1v3$ are almost 20% difference emphasizing the effect of accounting for the differences. Still, it is not possible to have a well-scaled data in every situation, thus, seeing that using *RAW* yields performances better than random hints at the potential the pupil might have even though there are individual differences in pupil dilation caused by the easy and the more difficult condition. The 0-calibration uses only ten seconds of data to compute the mean for scaling the proceeding trial run resulting in better performances than *RAW*. This might indicate that very small resting phases could be used for a continuous re-calibration in a real-time scenario. However, the scaling is applied to a close in time trial run which itself is not very long. Thus, it remains to be seen if such a calibration works for longer trial runs.

Accounting for differences with a rather simple methodology appears to work well for the pupil radius. Not using any form of scaling in contrast to using either *SCALED* or *CALIB* yields significantly better performance results as expected. Next, the effect of *FEATURES* is investigated.

5.5.4 Goal 4: Utilizing the pupil

Two sets of features were used for training *FEATURES*9 and *FEATURES*27. In the beginning, it was stated that the other *FACTORS* are ignored when analysing a single *FACTOR* . Other machine learning algorithms than the RFC might require other processing steps such as feature scaling and selection. Further, the number of trees might have an impact as ten random trees might not make use of all 27 features when randomly selecting the features for splitting while 50 trees might do as the chance is increased to select a feature.

The mean accuracies for each set and *SC* is depicted in figure 13, the appendix includes all statistics (see E.3.3).

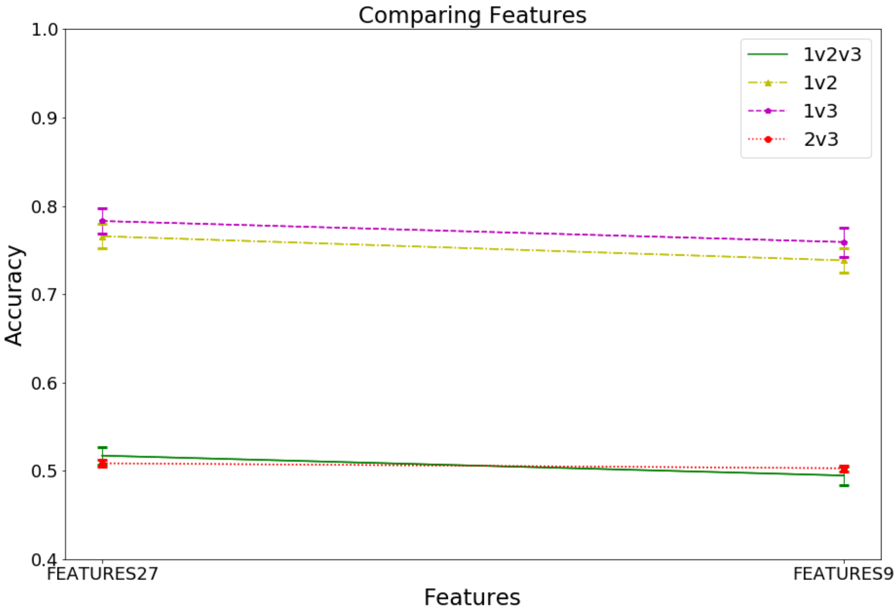


Fig. 13. Mean accuracies for each of the feature sets with standard deviation bars.

Using more features seems to affect the average performance positively, however, the difference, for instance, for $1v2$ is 2.8% indicating a small increase. For $2v3$ the difference is only 0.06%; thus, more features do not seem to improve performance when the differences of the pupil are too subtle.

Two sets do not require a Friedman test. Instead, Wilcoxon was used directly. The tests were significant except for comparing $2v3$. The number of samples compared was 30 per *FEATURES* .

From the results, it can be assumed that using the additional features from the derivative signal and the percentage change signal is useful when using the pupil radius and statistical features.

5.5.5 Goal 5: Validation

For each pipeline CV was used, thus, each accuracy score can be seen as generalizable. As the investigation of Goal 1-4 used averages of the CV scores generalization can be questioned as they were trained with the same data. The choice of non-parametric tests was informed by this challenge; however, the analysis does not account for interaction effects. Looking at the main goal of cross-user classification it can be stated, due to using the LOUOCV, the performance results show the feasibility of utilizing pupil radius in the context of the n-back task with letters for cross-user classification; thus, RQ1 can be answered positively. Whether all *FACTORS* effects are generalizable is of less importance, but the results indicate their influence.

With the results of Goals 2-4, a pipeline can be chosen that according to the analysis should yield the best results. This pipeline can then be used in the second study where new data is gathered. Therefore, the generalization can be taken one step further.

It was stated that comparing individual models with the population models can be used as a form of validation. It is assumed that classifying for the individual yields better performance results, due to the physiological differences. The individual models are discussed in the next subsection.

5.5.6 Individual Models

For each participant the 240 pipelines were used to create models and each pipeline used leave-one-trial-run-out CV. The population models were working well for $1v2$ and $1v3$ this is expected to hold for individual models as well. While the difference across all users for $2v3$ was not distinguishable, for the individual it might be. Table 10 shows the average performance scores for each individual over all pipelines and *SC* . Additionally, the average cross-validation scores when the participant was excluded from training are displayed to get an idea of how well the individual model performed in contrast to the population models.

Participant	1v2v3	1v2	1v3	2v3	1v2v3 CV	1v2 CV	1v3 CV	2v3 CV
P1	0.562	0.723	0.801	0.579	0.462	0.635	0.765	0.527
P2	0.721	0.838	0.877	0.721	0.489	0.758	0.771	0.475
P**3**	nan	0.795	nan	nan	0.486	0.71	0.742	0.498
P4	0.647	0.848	0.859	0.641	0.525	0.84	0.668	0.513
P5	0.764	0.891	0.941	0.744	0.415	0.655	0.586	0.516
P6	0.737	0.887	0.962	0.717	0.487	0.754	0.824	0.488
P8	0.624	0.877	0.84	0.629	0.51	0.866	0.786	0.418
P10	0.559	**0.793**	0.804	0.503	0.53	**0.878**	0.753	0.499
P12	0.536	**0.756**	0.775	0.507	0.546	**0.801**	0.731	0.519
P13	0.65	0.788	0.823	0.675	0.479	0.787	0.801	0.449
P15	**0.788**	0.886	0.915	**0.836**	0.613	0.823	0.901	0.576
P16	0.544	**0.619**	0.688	0.691	0.514	**0.785**	0.859	0.452
P17	0.596	0.833	0.878	0.534	0.459	0.669	0.733	0.508
P**19**	0.535	**0.653**	0.807	0.547	0.48	**0.673**	0.673	0.538
P20	0.556	0.887	0.943	0.401	0.528	0.728	0.813	0.57
P21	0.636	0.941	0.85	0.596	0.516	0.789	0.855	0.454
P22	0.561	0.712	0.885	0.638	0.451	0.626	0.657	0.543
P23	0.483	**0.662**	0.681	0.522	0.522	**0.761**	0.766	0.515
P31	0.592	0.713	0.814	0.668	0.55	0.867	0.895	0.455
P33	0.615	0.884	0.713	0.701	0.562	0.85	0.853	0.511
P35	0.607	0.738	0.871	0.68	0.556	0.657	0.836	0.608
P36	0.446	0.707	0.607	0.548	0.417	0.709	0.639	0.426
P39	0.711	0.844	0.868	0.746	0.493	0.709	0.771	0.487
P48	0.736	0.944	0.871	0.74	0.556	0.718	0.824	0.595
Mean	**0.618**	**0.801**	**0.829**	**0.633**	**0.506**	**0.752**	**0.771**	**0.506**
SD	0.089	0.091	0.087	0.101	0.046	0.076	0.081	0.049
Min	0.446	0.619	0.607	0.401	0.415	0.626	0.586	0.418
Max	0.788	0.944	0.962	0.836	0.613	0.878	0.901	0.608

Table 10. Descriptive statistics specificity individual models and the average validation score when the participant was left out during LOUOCV (indicated by CV). Bold highlights are discussed in the text

P3's 3-back data was dropped as mentioned in the analysis section due to noise; thus, those scores are missing. During population model training P3 could, however, be used as a test set in all four SC . On average the individual models perform slightly better compared to the population models despite having fewer data available for training. For P15 the models performed well even if 3-back was part of the classification (78% and 83%). For P10, P12, P16, P19 and P23 $1v2$ accuracies were better during LOUOCV this as well counts for some when looking at $1v3$. This might indicate the advantage of population models where more data is available to learn.

As a form of validation of the performance of the population models, individual models were inspected which performed slightly better on average.

5.5.7 Best Performing Pipelines

Several different pipelines were trained using cross-validation, average classification scores across all pipelines grouped by *SC* and by the different *FACTORS* . The averages included the expected to be poorly performing approaches that did not use scaling which was referred to as *RAW*. While the average across all the pipelines allowed to investigate the different goals. It is now interesting to see what accuracies can be reached. Because at the end only one pipeline is supposed to be used for building a model. In the following, it is discussed which of the pipelines performed best to get an idea of what is possible with the present approaches. Only the *SC* contrasting 1-back and either 2- or 3-back are considered as the others cannot be distinguished well enough.

In contrast to taking the average score of all pipelines, the most promising ones are briefly highlighted. With the investigation fo Goal, 1-4 approaches of interest are reduced to those using *CALIB* or *SCALED* representing the two approaches for accounting for individual differences. Further, only *RFC*50 is considered, and the window sizes *WIN*5, *WIN*10, *WIN*30, *WIN*60. Thus, table 11 shows the best performance for each window size, scaling approach, and two-class classification.

WIN	SC	$INDIV$	$FEATURES$	Accuracy	SD	Class Sizes	Confusion Matrix
$WIN60$	$1v2$	$SCALED$	$FEATURES27$	0.891	0.085	[95, 96]	[0.874 0.125] [0.095 0.906]
$WIN30$	$1v2$	$SCALED$	$FEATURES27$	0.86	0.094	[3319, 3347]	[0.851 0.148] [0.134 0.867]
$WIN10$	$1v2$	$SCALED$	$FEATURES27$	0.838	0.098	[5219, 5267]	[0.808 0.19] [0.135 0.866]
$WIN5$	$1v2$	$SCALED$	$FEATURES27$	0.826	0.094	[5694, 5747]	[0.795 0.203] [0.146 0.855]
$WIN60$	$1v2$	$CALIB$	$FEATURES27$	0.812	0.173	[95, 96]	[0.842 0.156] [0.221 0.781]
$WIN30$	$1v2$	$CALIB$	$FEATURES27$	0.771	0.143	[3319, 3347]	[0.847 0.152] [0.307 0.696]
$WIN10$	$1v2$	$CALIB$	$FEATURES27$	0.759	0.113	[5219, 5267]	[0.819 0.179] [0.302 0.7]
$WIN5$	$1v2$	$CALIB$	$FEATURES27$	0.739	0.112	[5694, 5747]	[0.781 0.217] [0.304 0.699]
$WIN60$	$1v3$	$SCALED$	$FEATURES27$	0.901	0.11	[95, 91]	[0.884 0.121] [0.074 0.923]
$WIN30$	$1v3$	$SCALED$	$FEATURES27$	0.867	0.121	[3319, 3209]	[0.866 0.138] [0.121 0.875]
$WIN10$	$1v3$	$SCALED$	$FEATURES27$	0.857	0.119	[5219, 5049]	[0.847 0.158] [0.124 0.871]
$WIN5$	$1v3$	$SCALED$	$FEATURES27$	0.844	0.122	[5694, 5509]	[0.831 0.175] [0.132 0.863]
$WIN60$	$1v3$	$CALIB$	$FEATURES9$	0.847	0.157	[95, 91]	[0.884 0.121] [0.189 0.802]
$WIN30$	$1v3$	$CALIB$	$FEATURES27$	0.847	0.109	[3319, 3209]	[0.88 0.124] [0.18 0.813]
$WIN10$	$1v3$	$CALIB$	$FEATURES27$	0.806	0.096	[5219, 5049]	[0.828 0.178] [0.205 0.788]
$WIN5$	$1v3$	$CALIB$	$FEATURES27$	0.791	0.098	[5694, 5509]	[0.819 0.188] [0.227 0.765]

Table 11. Performance of individual pipelines for: $1v2$ and $1v3$

The best scores were 89.1% for $1v2$ and 90.1% for $1v3$ using $WIN60$. These same pipelines might be used for a final model that can be used in the second study.

Looking back at the last section where the pupil data of the data gathering study was inspected it was noted that there seems to be a trend showing increased pupil dilation during the first trial runs of the difficulty levels. While a statistical analysis was not conducted to investigate the observation, to experiment, the complete first trial runs were excluded from the data, and the same 240 pipelines were trained. Table 12 shows the results of the best performing pipelines as presented in table 11.

WIN	SC	$INDIV$	$FEATURES$	Accuracy	SD	Class Sizes	Confusion Matrix
$WIN60$	$1v2$	$SCALED$	$FEATURES9$	0.944	0.094	[72, 72]	[0.958 0.042] [0.069 0.931]
$WIN30$	$1v2$	$SCALED$	$FEATURES27$	0.908	0.101	[2515, 2512]	[0.938 0.062] [0.121 0.879]
$WIN10$	$1v2$	$SCALED$	$FEATURES27$	0.88	0.095	[3955, 3952]	[0.88 0.12] [0.119 0.881]
$WIN5$	$1v2$	$SCALED$	$FEATURES27$	0.871	0.097	[4315, 4312]	[0.874 0.126] [0.133 0.867]
$WIN60$	$1v2$	$CALIB$	$FEATURES27$	0.799	0.177	[72, 72]	[0.833 0.167] [0.236 0.764]
$WIN30$	$1v2$	$CALIB$	$FEATURES27$	0.814	0.139	[2515, 2512]	[0.875 0.125] [0.248 0.752]
$WIN10$	$1v2$	$CALIB$	$FEATURES27$	0.786	0.127	[3955, 3952]	[0.848 0.152] [0.276 0.724]
$WIN5$	$1v2$	$CALIB$	$FEATURES27$	0.756	0.13	[4315, 4312]	[0.807 0.193] [0.295 0.705]
$WIN60$	$1v3$	$SCALED$	$FEATURES27$	0.972	0.08	[72, 68]	[0.986 0.015] [0.042 0.956]
$WIN30$	$1v3$	$SCALED$	$FEATURES27$	0.934	0.105	[2515, 2408]	[0.955 0.047] [0.082 0.914]
$WIN10$	$1v3$	$SCALED$	$FEATURES27$	0.905	0.109	[3955, 3788]	[0.919 0.085] [0.101 0.894]
$WIN5$	$1v3$	$SCALED$	$FEATURES27$	0.893	0.109	[4315, 4133]	[0.904 0.1] [0.111 0.885]
$WIN60$	$1v3$	$CALIB$	$FEATURES9$	0.875	0.123	[72, 68]	[0.917 0.088] [0.153 0.838]
$WIN30$	$1v3$	$CALIB$	$FEATURES9$	0.87	0.115	[2515, 2408]	[0.925 0.078] [0.178 0.814]
$WIN10$	$1v3$	$CALIB$	$FEATURES27$	0.823	0.118	[3955, 3788]	[0.864 0.142] [0.209 0.782]
$WIN5$	$1v3$	$CALIB$	$FEATURES27$	0.81	0.105	[4315, 4133]	[0.848 0.159] [0.217 0.773]

Table 12. Performance of individual pipelines for: $1v2$ and $1v3$ without including the first trial run

The best scores were 94.4% for $1v2$ and 97.2% for $1v3$ using $WIN60$. Accuracy increased without the first trial run indicating that there was some adjusting to the difficulty which influenced the pupil dilation and potentially the cognitive state. However, the design of the experiment did not account for this as the practice phase was supposed to let participants develop a strategy and get familiar with it.

High accuracies have been reached for the two-class classification presented in table 11 and 12. The next section will briefly summarize the results of classifying the cognitive state across users.

5.6 Summary

This section addressed RQ1 and partially was able to answer it positively. The classification input was constrained to pupil data with which population models were built able to discriminate $1v2$ and $1v3$ reasonably well (75%, 77%: average over all pipelines), the others could not be distinguished most likely as a result of the insignificant differences between 2-back and 3-back. Thus, by simply looking at Goal 1 the first research question could already be answered. In order to investigate the potential influences of $FACTORS$ on performance, Goals 2-4 were formulated. Bigger window size, $FEATURES27$, $CALIB$, $SCALED$ were found to affect performance positively. Including all data, up to 90% accuracy was reached showing the feasibility of the present approaches.

For the main research goal, concerned with cross-user and cross-task classification of the cognitive state using physiological measures, one step was taken. The data gathering study was conducted and used for machine learning. For the latter, the results were presented in this section. The building of models working across users, thus, the question asked by RQ1 can be answered: using pupil dilation labelled with the difficulty levels - as representatives of the cognitive state - of the n-back task can be used to build models across users for $1v2$ and $1v3$. The results allow to take on RQ2 concerned with cross-task classification. Subsequently, a second study was conducted which is subject of the next section.

6 CLASSIFYING COGNITIVE STATE ACROSS TASKS

The first study collected pupil data from 24 participants which was used to train models that were able to distinguish between the 1-back and 2-back condition reasonably good. As a result, it is possible to continue with the main research goal of doing cross-user and cross-task classification. While it was successful to classify the n-back task using letters using population models the second research question RQ2 is concerned with cross-task classification. In order to address research question 2 (RQ2), the second study required participants to perform four variants of the n-back task.

First, the second study is described shortly followed by an analysis and presentation of the results similar to section 4.2. Then the newly gathered data is used for machine learning, and the results are highlighted.

6.1 Experiment

Design, procedure, setting, task and apparatus all were very similar to the data gathering study to keep the same structure of the experiment. The differences and adjustments made for the second study are highlighted, thus, omitting redundant information presented in section 4.1. In contrast to the data gathering study, only two difficulty levels were performed by the participants (1-back and 2-back) as the models built in the last section had the results for these two levels due to the issue of having no significant difference between the 2-back and 3-back conditions. The models were trained using pupil radius only; thus, the E4 was not used in the second study.

6.1.1 N-back Task Variations

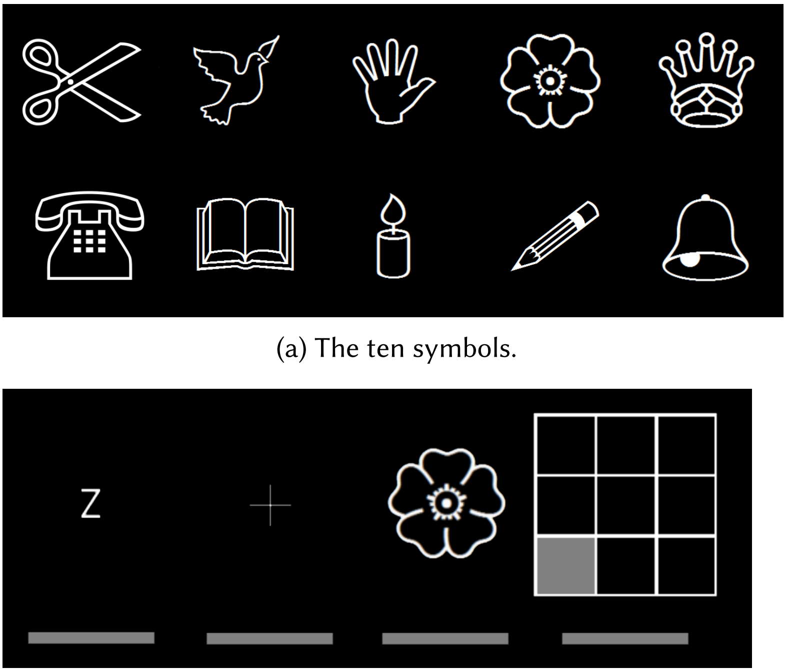
The study prototype was extended with three more variants of the n-back task. The first two are based on Grimes et al. [31]. In one variation instead of letters symbols are presented another displayed a 3x3 grid where only one element of the grid is highlighted as stimulus. Additionally, an audio version of the letter task was added. The structure of the task is, therefore, the same as with letters only the stimuli differ.

Visual letter n-back. The task of the data gathering study was used without making any adjustments.

Audio letter n-back. The audio task used the same letters as the letter task. They were recorded by a female singing teacher with clear pronunciation. The eye tracker requires the participants to keep their eyes open. Thus, a cross-hair was displayed in the middle of the screen (see figure 14b) which they were told to focus on. During the other conditions, they always had to look at the same position where the stimuli were presented.

Visual symbol n-back. The symbol task took a subset of the symbols used in a spatial memory study [79]. All ten can be seen in figure 14a. All have a white silhouette and if they are written down the words do not contain more than two syllables in the German language.

Visual spatial n-back. The spatial task used a 3x3 grid instead of a letter or symbol they had to remember the location of the element that was highlighted on the grid. In contrast to the other tasks, only nine distinct stimuli are given for the 3x3 grid. A question was how big the grid should be as it potentially is harder to remember one of 16 locations in a 4x4 grid. At the end, as in [31] a 3x3 grid was used.



(a) The ten symbols.

(b) The four stimuli types in the way they were presented in size with the feedback bar below as reference. Left to right: letter, audio, symbol, spatial.

Fig. 14. The four n-back variations

6.1.2 Adjusted Design & Procedure

All participants started with the letter n-back task and then performed the three other tasks in a counterbalanced order. It was the same *latin squares* order which

was used for the task with three difficulty levels previously. Each task variation was played with two difficulty levels (n = 1, 2). They always started with 1-back for every task type as done by Grimes et al. [31] who state that this would be more representative of a scenario where the user first has to generate training data before using the real system. Each difficulty level had three trial runs instead of four. The total amount of trial runs participants had to perform ($4\ tasks * 2\ levels * 3\ trialruns = 24\ trial\ runs$), therefore, was doubled in contrast to data gathering study ($1\ task * 3\ levels * 4\ trial\ runs = 12\ trial\ runs$). After each difficulty level, the NASA-TLX was filled in. An extra break of one minute was added after both levels of a variation were completed. Before each, an explanation of the new variant was presented along with a practice phase. The practice phase of the letter n-back task was longer in contrast to the others as it was expected that due to the similarity of the task not much training would be required.
There was no compensation for participating. In the post-interview, the participants were additionally asked to explain which variations they felt most comfortable with and how they remembered the stimuli.

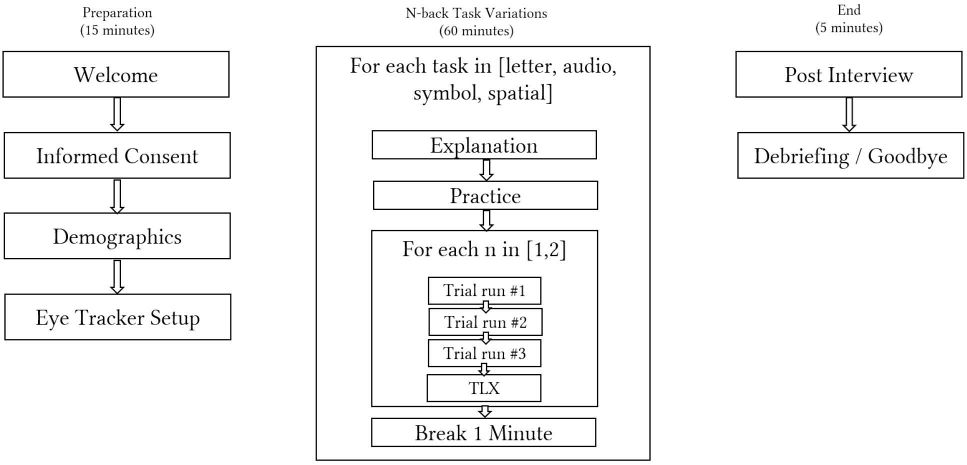


Fig. 15. Adjusted procedure of the second study.

6.1.3 Pilot Study

A pilot study was conducted to make sure the new tasks are working as expected. Also, to get an idea if the increased duration was too much potentially causing noticeable fatigue effects. Further, the audio presentation had to be checked whether the letters were understandable.

One female participant took part in the pilot study. From a technical perspective, everything went fine, and all tasks worked as expected as well as the data recording. She stated that C and Z sounded a bit similar, but they were kept. The new task explanations were understood and the practice time for each variation was perceived as sufficient. The participant stated that focusing in the middle of the screen for so long is exhausting for the eyes. As a result, the one-minute break was added afterwards, so participants were able to relax their eyes if required. In summary, the pilot study did not reveal any significant issues.

6.1.4 Summary

The second study was presented in this section with a focus on the differences compared to the data gathering study. Before the newly gathered data is used for machine learning, first, a typical analysis of the experiment is presented next.

6.2 Results & Analysis

For the second study six (three male, three female) participants were recruited 23 to 28 years old. All were students of the computer science department. Three participants rated their cognitive handicap with three or lower, and the rest ranged between 5 and 8.

The four variants will be referenced as follows. *LETTER* for the n-back task with letters, *AUDIO* for the variant presenting the letters via audio, *SPATIAL* for the spatial n-back variant with the grid and *IMG* for the task with the symbols. Due to the small amount of participants only descriptive statistics are presented, and no inferential tests were done. Further, the descriptive statistics in the case of presenting averages should be taken with caution. It is expected that the tendency of the performance and NASA-TXL ratings are in line with the expected difficulty levels. Further, for the pupil, only an inspection of the signals for each task is given to see if the same tendency can be observed as in the more extensive data gathering study.

6.2.1 NASA-TLX, Performance, Post Interviews

The results of the TLX ratings are presented in table 13. The general tendency of 1-back ratings being on average smaller than 2-back is the same as in the explorative study. However, the range is quite large as the ratings between subjects differed. The ratings of *LETTER*and *AUDIO* tend to be lower on average in contrast to the ratings of *IMG* and *SPATIAL*.

	LETTER		IMG		SPATIAL		AUDIO	
	1-back	2-back	1-back	2-back	1-back	2-back	1-back	2-back
P49	63.333	67.5	75.0	80.0	50.833	45.0	61.667	59.167
P50	38.333	51.667	45.0	61.0	36.667	58.333	50.0	63.333
P51	52.5	55.0	60.0	62.0	44.167	65.0	50.833	68.333
P52	29.167	41.667	34.0	49.0	26.667	30.833	31.667	35.0
P53	50.0	53.333	59.0	63.0	51.667	48.333	57.5	40.833
P54	36.667	26.667	42.0	30.0	22.5	23.333	27.5	35.833
Mean	45.000	49.306	38.750	45.139	40.278	47.083	46.528	50.417
Median	44.167	52.500	40.417	46.667	44.167	50.417	50.417	50.000
Std. Deviation	12.506	13.828	12.301	15.868	18.534	18.898	13.879	14.877
Range (Max - Min)	34.167	40.833	29.167	41.667	53.333	43.333	34.167	33.333

Table 13. NASA-TLX ratings for the four n-back task variants.

Further, the rating for 2-back is in some cases lower than for 1-back. For instance, P49 (*SPATIAL*) or P53 (*AUDIO*). An explanation could be that there is a form of accommodating to the new variant, thus, in the 1-back condition, more effort is subjectively put in.

The performance (see appendix F), measured in terms of accuracy and response time, tended to be marginally worse in the 2-back condition, meaning the response times were longer for 2-back and accuracy decreased. This can be observed for each variant.

For the post interviews, an additional question was added to get an idea which of how the task variants difficulties were perceived. Due to the removal of 3-back, it was not a surprise that participants did not have to guess much. P52 sometimes forgot the stimulus of the 0-back task. P54 had to guess intuitively during the audio variant and rarely during the 2-back conditions of the other. P54's cognitive handicap rating was 8; hence, the exhaustion of the participant might be a reason for not being very concentrated. Three stated that there was a big jump in difficulty between 1-back and 2-back conditions. The latter fits the TLX ratings, and it can be assumed that the performance would reflect this as well with more participants.

Participants were asked to rate the four variants from easiest to hardest. As five said that two variants were equally hard for them, three instead of four difficulty levels were used, thus, easy, medium and hard. The results of the frequencies a task was rated as easiest, middle or hardest can be seen in table 14. The spatial and symbol task were easier for the participants in contrast to the letter and audio variant. The latter was rated three times as hardest. An explanation for this might be that they had to keep their eyes open and focus in the centre of the screen (P49, P52). P54 stated that he would have wished to be able to close his eyes. P50 found that the pronunciation of the letters was weird for her. P52 wanted to repeat the letters in his head which interfered with the audio output.

	LETTER	IMG	SPATIAL	AUDIO
Easiest	0	3	4	1
Middle	4	3	1	2
Hardest	2	0	1	3
Overall	14	9	9	14

Table 14. Frequencies of perceived difficulty of the four task variants. The overall score was computed by multiplying the frequency with 1,2 or 3 for easiest, middle, hardest.

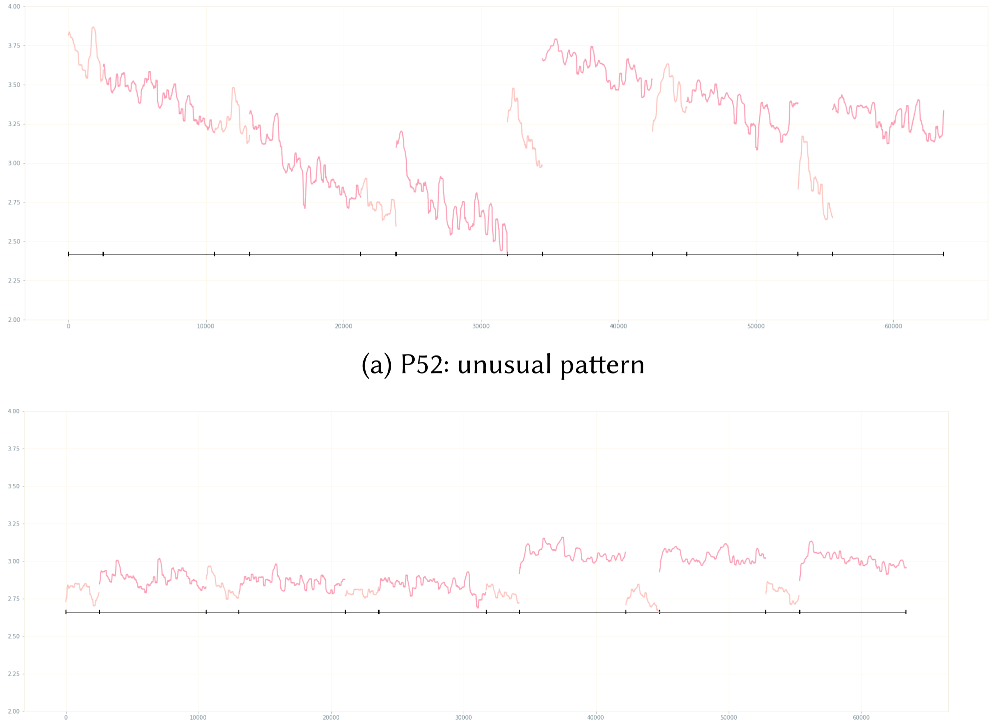
Several made a statement regarding how they approached the task. For *LETTER* three explicitly stated that they repeated the sequence loud in their heads. For *SPATIAL* one remembered the position by storing "upper left", hence, verbally. Two stated that they remembered the path (2-back) of the stimulus. For 1-back it was very easy for them as they only had to stare at the same position in the grid. In *IMG* two stated they tried to remember the symbols visually while others used words in the 2-back condition.

The analysis of the questionnaires and interviews indicated that the difficulties of each task variant are sufficiently different. Hence, it can be expected that the pupil will be sensitive to these changes as it was in the data gathering study. The interviews with the participants indicated they had different approaches to storing the information (e.g. grid locations). Therefore, it cannot be said for sure whether the difficulty levels of the four variants are all equal because of the between task variances and individual differences. The next section is going to take a look at the pupil data.

6.2.2 Pupil Analysis

As done in the data gathering study the pupil data was inspected for each participant to see which data might be too noisy. For P51 and P54 the signal was very good and least noisy. The others were noisy but could be smoothed with a sufficiently good outcome.

It was further inspected whether the pattern 1-back pupil size is smaller compared to 2-back for each variant can be observed by plotting the data. Only P52 had a slightly different pattern pupil data continuously decreased during each condition (see figure 16). In the post-interview, P52 stated that he was not very concentrated and tired. That might have caused the pattern of decreasing diameter over time which can be interpreted as an increased effort put in by the participant during the first trial run which could not be kept during the whole condition indicating a fatigue effect.



(a) P52: unusual pattern

(b) P50: expected pattern

Fig. 16. P50’s and P52’s 1-back and 2-back data for *LETTER*. The brighter colour highlights the 0-back phase. The darker colour highlights the n-back phases.

	LETTER		IMG		SPATIAL		AUDIO	
	1-back	2-back	1-back	2-back	1-back	2-back	1-back	2-back
P49	2.936	3.232	2.767	2.998	2.657	2.694	2.625	2.705
P50	2.864	3.031	2.747	2.934	2.674	2.855	2.581	2.848
P51	3.19	3.432	3.23	3.479	2.954	3.202	2.942	3.143
P52	3.049	3.412	2.841	3.145	2.763	2.808	2.468	2.69
P53	2.918	3.081	2.748	2.81	2.625	2.902	2.719	2.804
P54	3.102	3.628	3.105	3.537	2.909	3.489	3.202	3.584
Mean	3.010	3.303	2.906	3.150	2.764	2.992	2.756	2.962
Median	2.992	3.322	2.804	3.071	2.719	2.879	2.672	2.826
Std. Deviation	0.125	0.229	0.209	0.298	0.139	0.297	0.271	0.346
Minimum	2.864	3.031	2.747	2.810	2.625	2.694	2.468	2.690
Maximum	3.190	3.628	3.230	3.537	2.954	3.489	3.202	3.584

Table 15. Mean pupil radius of all four variants for each participant

Table 15 shows the descriptive statistics of each participant and across the different variants. The mean values indicate a difference between 1-back and 2-back for each variant. The difference is, however, for some participants very small. For instance, P49's (*SPATIAL*) difference is 0.037 millimetres in contrast to P54 with 0.58 mm. The pupil seems to be sensitive to the variations, yet in some cases the difference is small. In order for the classifier to work the difference between the two conditions needs to be similar in magnitude to the data of the first study. To investigate this differences further each participant's data was visually inspected (as done for noise). The result can be seen in Table 16.

	LETTER	IMG	AUDIO	SPATIAL
P49	acceptable	good	bad	bad
P50	very good	very good	good/very good	very good
P51	good/very good	acceptable	good	good
P52	good	good	good	bad
P53	acceptable	acceptable/bad	very good	acceptable
P54	very good	very good	very good	very good

Table 16. Manual inspection of the difference between 1-back and 2-back for each task variation.

Explanation:

bad: values almost in the same range

acceptable: difference visible, but lots of overlap

good: some overlap but clear difference visible

very good: clear difference no overlap or only marginally

Each pair of 1-back and 2-back values were compared. If the values for each condition are from different value ranges, good discrimination between both can be expected. For example figure 16 shows *LETTER* for P50 and P52. P50 has a *very good* visible difference while there is some overlap for P52 - *good* - as at the beginning of 1-back the pupil size is comparable to the 2-back pupil values.

Also, there seem to be differences between the variants. Ideally, the mean for a participant during 1-back is the same across all variants; however, each variant might require different effort by the participant leading to differences.

6.2.3 Summary

The short description and analysis of the second study during which pupil data for four variants of the n-back task was gathered were presented. As expected NASA-TLX scores and pupil radius were different when comparing 1-back with 2-back, however, in some cases this differences is small. Next, the data is used to make classifications not only across users but also task variants.

6.3 Classifying Cognitive State: Across Task and Users

In order to address research question 2 (RQ2), the second study, requiring participants to perform four variants of the n-back task, was conducted. The data of the six participants can be used to test the model of the data gathering study and will indicate how it performs with new data from the same and different task variants.

6.3.1 Preprocessing, Choice of Population Model, Data Scaling

Section 5.5.7 discussed the best performing pipelines to train a final model. For the classification across variants of the task only four pipelines used for classification are reported, namely, those using *RFC*50, *FEATURES*27, *WIN*10 or *WIN*30 and either *CALIB* or *SCALED*. As for the 60-second windows only three samples per difficulty - one for each trial run - would be available it was left out. Smaller windows performed worse than larger windows. Still, 10 and 30 seconds are suitable choices for a real-time or evaluation oriented system.

The population model can be used to discriminate between the two difficulty levels of each variant. It cannot be assumed that each difficulty condition is eliciting the same pupil response across the variants. Reasons for this could be the stimuli type or individual preferences. Therefore, to be able to answer the question if the model can be used across task variants there is a need to account for the differences between them. The population model accounts for individual differences by scaling the data of each participant. To account for differences between task variants their data is scaled separately. Instead of using *SCALED*, the reference *SPT* short for scaled per task is used. As noted previously such scaling is not possible in a real-time scenario where not all data is available. *CALIB* is an attempt only applicable if short calibration phases during real-time interaction can be integrated. An alternative is to have a training phase in the beginning before interaction. To simulate this the data of the n-back letter variant which always proceeded the other variants in the second study could be used to scale the four variations' data. However, this does not allow to account for the differences between the variations which makes it hard to evaluate the classification performance. The output of a classifier would be relative to the *LETTER* condition. Therefore, the approach is not reported. A possibility for a real-time approach would be to use the data of the population model for scaling. This comes with the same issue of not being able to account for the task variation differences.

All data gathered in the first study was used to build models while the data of six new participants was used to test the models with each variant of the task. The data of the new participants was processed the same way as those of the first study except for using an additional scaling approach besides *CALIB*, namely, *SPT*.

6.3.2 Results: *CALIB* and *SPT*

Table 17 shows the classification accuracy of each participant for each pipeline.

WIN10	LETTER		IMG		AUDIO		SPATIAL	
	CALIB	*SPT*	*CALIB*	*SPT*	*CALIB*	*SPT*	*CALIB*	*SPT*
P49	0.581	0.863	0.724	0.773	0.706	0.612	0.597	0.588
P50	0.588	0.927	0.958	0.997	0.758	0.933	0.761	0.906
P51	0.67	0.936	0.748	0.827	0.826	0.927	0.594	0.758
P52	0.851	0.726	0.574	0.739	0.852	0.785	0.561	0.53
P53	0.524	0.758	0.506	0.712	0.809	0.936	0.591	0.794
P54	0.775	0.994	0.997	0.988	0.93	0.976	0.903	1.0
Mean	0.665	0.867	0.751	0.839	0.814	0.862	0.668	0.763
WIN30	LETTER		IMG		AUDIO		SPATIAL	
	CALIB	*SPT*	*CALIB*	*SPT*	*CALIB*	*SPT*	*CALIB*	*SPT*
P49	0.627	0.895	0.724	0.848	0.738	0.776	0.652	0.71
P50	0.586	1.0	0.976	1.0	0.648	0.976	0.605	0.914
P51	0.589	1.0	0.746	0.99	0.87	1.0	0.543	0.829
P52	0.837	0.833	0.569	0.775	0.914	0.933	0.586	0.386
P53	0.529	0.776	0.471	0.79	0.938	1.0	0.595	0.924
P54	0.77	1.0	1.0	1.0	0.986	1.0	1.0	1.0
Mean	0.656	0.917	0.748	0.901	0.849	0.947	0.664	0.794

Table 17. Accuracy scores of cross-task classification for each participant and task variant. Top with *WIN*10, bottom with *WIN*30.

The population model was able to discriminate between the difficulty conditions well when *SPT* and *WIN*10 was used. *LETTER* and *AUDIO* reach 86.7% and 86.2% accuracy on average, *IMG* 83.9% and *SPATIAL* 76.3%. With larger windows of 30 seconds, accuracies range between 79.4 % and 94.7% improving the average accuracy. Looking at the individual scores *WIN*30 increased the performance by more than 10%, for instance, for P49 (*AUDIO*) the score went up 16.4%. Recalling the inspection of the data concerned with the difference between 1-back and 2-back (table 16) the difference for the mentioned condition of P49 was rated as *bad*. It seems the window size might be useful in such cases where both conditions produce similar pupil responses. In the case of P52 (*SPATIAL*) the performance decreased by 14% indicating that window size does not always help. Larger windows captured the general trend of the pupil size better. For instance, the 1-back pattern of P52 seen in figure 16 decreases over time and is higher at the beginning of a trial run. A window cannot account for the decrease over time if the decrease is not visible in the range of the window. Thus, a smaller window performs worse in the case of P52.

While the *SPT* approach yields good results, *CALIB* only performs well in some cases. *IMG* and *AUDIO* reach 75.1% and 81.4% on average using *WIN*10. *LETTER*, as well as *SPATIAL*, show mediocre classification performance, 66.5% and 66.8%. A larger window only affects the performance of *AUDIO* positively. *CALIB* depends on the 0-back condition to elicit the same pupil pattern independent of difficulty level (see P50 in figure 16 where this is the case). Still, classification performance is only 58.8% because the 1-back condition is classified as 2-back in most cases. Thus, what the model learned to be 1-back is not the same as the pattern P50 shows. Although there is a clear difference between both conditions. Using *SPT* allows scaling of the data in a way that even if the 1-back of the new data is different, it can classify it with better accuracy.

RQ2 asked whether the cross-task classification is possible: the results indicate that it is. However, the testing with six participants only gives one single score for each participant and pipeline; thus, generalization should be made with caution.

6.4 Summary

A small study to collect data for four variants was conducted, performance and TLX scores indicated that for each variant both difficulty levels were different, thus, potentially producing different states.

The results for cross n-back variant classification are very promising. The results are further discussed in the next section which summarizes the work, its contribution, and provides a discussion of findings and contrasts those with related literature.

7 SUMMARY & DISCUSSION

In the following, a summary is given highlighting the main findings. Then, the contribution of the thesis regarding cognitive state classification is pointed out. While each section already discussed the results to a certain extent, this section continues with an additional discourse about noteworthy findings. Then, the most relevant related work is contrasted with the present research.

7.1 Summary & Contribution

This work investigated the classification of cognitive state across individuals (RQ1) and tasks (RQ2) using physiological measures to asses the state. The n-back task was utilized to modulate the state. The data gathering study collected pupil data of 24 users which was successfully used to build models across users. Unfortunately, the other measures (EDA, BVP) were not sensitive to the task's manipulation. The effect of window size, scaling approach and features were investigated showing that larger windows yield better performance results which as well holds for scaling the data per participant and using statistical features generated from the main pupil, the derivative and the percentage change signal. The models were not able to discriminate between 2-back and 3-back sufficiently well. Comparing 1-back with the two other conditions yielded reasonably good results with up to 90% accuracy. Contrasting the results of the population and individual models showed that the cross-user approach could not reach the performance of individual models. Subsequently, the second study gathered pupil data for four variants of the n-back task - letter, audio, spatial and image - of six participants. Scaling each task separately allowed the model to classify the difficulty levels of each variant with accuracies up to 94%.

Related work that as well used single-task settings along with working memory tasks to classify the cognitive state by the use of physiological measures mostly focused on the rather obtrusive EEG and individual models. Thus, this work contributes to cognitive state classification in the following ways:

- It was shown that pupil dilation could be used for cross-user classification. Furthermore, factors influencing performance were highlighted. Future work can make use of these findings when utilizing the pupil dilation for classifying cognitive state.
- Indications were found that a model built with pupil data of one task can be used to classify for other tasks.
- Eye tracking as a more unobtrusive and robust measure in contrast to EEG yields promising results helping researchers who would like to measure the cognitive state outside of the laboratory.

In the following, the findings and results presented in this work are discussed.

7.2 Sensitivity of EDA & BVP Measures

The data gathering study was conducted with the assumption that the three difficulty levels of the n-back task elicit different cognitive states. Those were measured with three physiological responses. Pupil dilation recorded with the mobile eye-tracking glasses, EDA and BVP with the E4 wristband. The latter's measures, however, did not prove to be sensitive to the task. Due to the general challenge of physiological differences between individuals, this result is not surprising. Moreover, similar scenarios where the measures did not show a response to the manipulation occurred in related work [64]. In some related work, it as well is the case that the physiological data is not analysed with inferential statistics before using it for building models [26] as they built individual models. Thus, insignificant results might not have been reported as data was directly fed to machine learning algorithms. The present data gathering study's analysis was conducted by comparing the data of three levels of difficulty at a very coarse level in contrast to the machine learning approach where, using sliding windows, a short segment (e.g. 10 seconds) of data in 1-back was assumed to be different from 2-back. The physiological data excluded could have been used as additional features. Feature selection similar to, for instance, [26], [70] or [31] could have been applied to find any useful responses for the E4's EDA and BVP measures. In their scenarios, individual models were under investigation. It is therefore concluded that feature selection for the individual to increase performance seems to be valuable. Nonetheless, in the present population model scenario, there is a need for features working across all users which lead to the decision of not including the insignificant measures for machine learning. Section 4.2.9 included thoughts on the insensitivity which are recapped and extended. Malik [43] suggests using at least two minutes of data when analysing heart rate measures. For each participant, there are four minutes of data per difficulty condition which most likely were not enough to observe significant changes. For instance, Solovey et al. [70] had roughly 48 minutes of data per participant successfully building individual models utilizing heart rate recorded with an ECG. Therefore, not only the amount of data was different but as well the measurement technique - ECG in contrast to BVP. Both might explain the insignificant results of the data gathering study. All of this was known beforehand. Nonetheless, more was expected due to recent work by Zhou et al. [80] to use peak measures of the raw BVP signal. The latter, however, did not show significant differences between conditions either. The author is not aware if there is work replicating the results by Zhou et al. [80]. Therefore, it might be assumed that their findings using a pipeline prediction task do not generalize. The E4 does measure the BVP at the wrist in contrast to their work where the sensor was attached to the fingertip. This might have an effect as well. While several arguments for the

BVP signal can be found, the EDA results are harder to explain. The signal was measured at the fingertips and with minimal noise which excludes one possible reason. The data was normalized per participant, thereby scaling the data to the same ranges for each participant to account for individual differences. A clear baseline measure approach might be more suitable for using the EDA signal. In [81] every measurement point was scaled based on the accumulated preceding trials which might have led to visible short-term changes if adopted to the present study. With more effort put into analysing the EDA signal changes if there are any might would have become observable which limits this work.

Both BVP and EDA measures might need more thorough analysis and more attempts to get the most out of the signal in contrast to the pupil dilation which is out-of-the-box useful. Still, it is not uncommon that a measure does not respond to the experimental manipulation.

7.3 2-back, 3-back & the Pupil

The analysis did not show significant differences between the 2-back and 3-back conditions. However, the NASA-TLX results, the performance measures and the post-interviews give a clear indication of a difference between both conditions. While the dilation was sensitive to the additional load indicated by the clear difference to 1-back, it might not be able to grasp the difference between the medium and high difficulty levels as it may only be able to discriminate between low load and elevated load. In [3], where pupil dilation was used, unfortunately, no 3-back condition existed to which the data gathering study could be compared. Others using EEG and a third difficulty level (e.g. [28], [31]) were able to discriminate between classes including 2- and 3-back. Therefore, EEG might be more specific in its discriminative power. In contrast to these two studies, the present n-back letter task was slightly different concerning timing modality but also regarding user feedback. The latter allowed participants to see, during 3-back, if their answers were correct which influenced their decision for the next letter, startled them if incorrect, and might have reduced the difficulty to a level which is similar to 2-back. This was not visible in the questionnaire ratings for which the explanation could be that the task appeared to be harder because of requiring to store one more letter, which was too hard for most participants and thus led to a physiological reaction not representing the difficulty but some form of excessive demand - not overload. With more practice and the possibility for each user to develop an approach suiting them to be able to deal with the 3-back condition, a different physiological response might have been elicited. Furthermore, with the individual models presented in section 5.5.6 it can be said that not being able to discriminate was not an issue of individual differences. Especially with the goal in mind to use the data of the first study to built a general model working across task and participant, a much longer training phase might have been the better approach. The outstanding classification results achieved by Gevins et al. [28] were built with data of participants practising the task for two days for six to eight hours before collecting data to build models. Similarly, Wilson and Russell [78] had a total of six hours of practice distributed across three days. Thus, it might be assumed that more practice would let users develop an approach for each difficulty level of the n-back task, and therefore, would not interfere with the data gathering by potentially elicit cognitive states that the task is not supposed to represent. In the present study, learning effects were accounted for by counterbalancing the difficulty levels. The learning effects within each difficulty level were not considered - a leverage point for future work to improve data gathering.

7.4 Research Question 1: Cross-User Classification

Two-class classification contrasting low (1-back) and elevated load (2-back or 3-back) performed reasonably well, reaching - averaged over all 240 pipelines - 75% ($1v2$) and 77% ($1v3$). The averages as well contain the results when no scaling or calibration was performed which does not account for differences among individuals - *RAW*. However, without scaling it is possible to get an idea of how bad the performance might get giving a lower bound. For *RAW*, using a window size containing the data of one complete trial run (60+ seconds) in combination with using all features led to accuracies of 79.1% ($1v2$) and 79.5% ($1v3$). With smaller window sizes, the performance ranged between 61.3% and 70.07%. Thus, if it is possible to reach an accuracy of 79%, for example, without accounting for subject-to-subject differences, it might indicate that the pupil dilation is not varying that much between subjects as to make such classifications possible without calibration. This can be seen as an advantage over other physiological recordings such as EDA which can be very different from one individual to the next. Moreover, EEG might be a more specific measure in its power to discriminate between multiple states but can be quite different as it captures even subtle differences. Therefore, using pupil measures might be more suitable in cross-user classification scenarios while EEG might be used for individual models. Looking at the other side - the upper bound so to say - best results were achieved for pipelines using scaling, in particular, using *WIN*60, *SCALED*, *FEATURES*27 and *RFC*50 led to 89.1% ($1v2$) and 90.1% ($1v3$). The other pipelines' performance results ranged from 79.1% to 88.5%. It can be seen that the maximum reached for *RAW* is equal to the worst performance using *SCALED*.

Discrimination between 2-back and 3-back or between all three conditions did not show good performance results. While the inferential analysis, as stated already, focused on comparing statistical features computed on the data of each complete difficulty condition, the sliding window approach in combination with the RFC was expected to might be able to find characteristics in the data that would allow for a discrimination even without significant results of the statistical analysis. The best result for $1v2v3$ was reached using 30-second windows, all features and scaled data - 62%. However, the accuracy only raised above a level of 50% because of the ability to discriminate the 1-back samples from the other two. Even though for $2v3$ 58% were reached using $WIN60$, all features and unscaled data which still appears random.

In a scenario where a solution to the given unsatisfactory performance is desired, the performance measures which were significantly different might be included in the predictive model. Thereby, information to be potentially able to discriminate between the two higher difficulty levels is added. Further, if the other physiological measures would be added, some improvement as well can be expected. However, with the given EDA and BVP data the chances are low. For a specific task which is supposed to be classified adding performance is a viable option. Yet, if the model is supposed to be general and applicable to other tasks, there needs to be a comparable measure that can serve as a performance indicator. In the case of working memory tasks performance is easily assessed which might not be the case for an arbitrary task.

Different factors influencing the performance of the population model were discussed and investigated. The tendencies affecting performance positive were found. In the end, only one set of parameters and choices of factors are supposed to be used for building a model that can be deployed in an interactive system or as an evaluation tool. Therefore, in the following, it is discussed what should be considered for systems that potentially use the gathered pupil data during the n-back task and the resulting predictive model.

First, regarding specificity, including 2-back and 3-back is not suitable due to not being significantly different. A system should discriminate between a low cognitive load level represented by the 1-back condition and a medium-to-high or elevated loaded level represented either by the 2-back condition or 3-back condition.
Second, the window size should be chosen too small as it affects performance negatively. However, a too big window might not grasp all short-term changes, but it can grasp the general tendency of the data better. With sliding windows, an output every second is possible. A question for interactive systems is how they are going to adapt to that output. For an evaluation tool where the real-time classification is not as crucial as for an interactive system larger windows might be more suitable.
Third, the accounting for individual differences can be done by scaling the data of each participant either using the 0-back approach or scaling over all data. In a system where real-time physiological data is recorded a comparable scaling needs to be applied in order for the classifier to work. In an evaluation scenario, the 0-back

approach is suitable, in an interactive system. However, it might be disturbing for the user to be asked for multiple calibration phases. Alternatively, only one calibration phase to scale the data might be performed for each user. Scaling the complete participant data before training a model does not involve a calibration phase. While it is possible to have the information for scaling during training a model, it is not available for real-time interaction and, as well, would require a training phase where data for scaling is gathered.
Fourth, when using the RFC, too many features are not an issue. According to the analysis, it seems that using statistical features of the derivative and percentage change signal slightly improves accuracy.

Overall, discrimination between low cognitive load represented by 1-back and elevated load represented by 2- and 3-back is possible which is why RQ1 could be answered successfully.

7.5 Research Question 2: Cross-Task Classification

Cross-task classification results were quite good using *SPT* and reached, averaged over all six participants, 86.7% for *LETTER*, 83.9% for *IMG*, 86.2% for *AUDIO* and 76.3% for *SPATIAL*, using all features and a window size of ten. *WIN*30 improved the performance up to 94.7% (*AUDIO*).

By applying *SPT* each variant is treated independently of the other, as long as a clear difference, regarding pupil dilation, between the difficulty conditions is present, the population model can discriminate between them well. It can be assumed that for each variant 1-back represents low working memory load and 2-back represents increased load; however, it cannot be assumed that across the variants every 1-back condition puts the same load on the user. Thus, each condition might represent a slightly different cognitive state. These between variants differences were removed *SPT*. The results, therefore, can only be interpreted as the population model is able to discriminate between low and elevated levels of load relative to a task variant.

The *CALIB* approach uses a baseline (0-back) to scale the data, which allows accounting for the differences to some extent as well but not as good as *SPT*: 66.5% for *LETTER*, 75.1% for *IMG*, 81.4% for *AUDIO* and 66.8% for *SPATIAL*. This approach heavily depends on the pupil response during the 0-back phase. For some individuals, it did not elicit the response which as expected. Hence, the results were rather mediocre. A continuous baseline was intended to be able to account for fatigue effects or unusual patterns which in the case of P52 (see figure 16) worked well (85.1%, *WIN*10).

A question one might ask is what the classifier is classifying. It is classifying difficulty relative to the variant. Thus, it is classifying cognitive state. If 1-back and 2-back used in the first study are supposed to represent low and elevated working memory load, the classifier can as well be used to classify each variant relative to these two levels. In that case, the between variant differences have to be kept. This way of classification could be used in an evaluation scenario where it is desired to assess the state of the user across several experimental conditions and have an additional or alternative measure to workload besides, e.g. the NASA-TLX.

7.6 Related Work

Section 2.3 introduced some relevant work attempting to assess the cognitive state. In the following, the results of this work are compared to those. There are many small differences one needs to be aware of when comparing classification performance results of other studies with the present results. Such as the window and step size they use for the sliding windows, the measure and the amount of data. Subsequently, the present results are contrasted with a subset of the discussed work in section 2.3 as a comparison requires to highlight subtle and large differences of the studies. Therefore, to keep it brief the work by Grimes et al. [31] which heavily inspired the ideas behind this thesis, the work by Appel et al. [3] which is one of the scarce examples using pupil dilation, n-back and cross-user classification, the work by Gevins et al. [28] as they are one of the few examples performing cross-user and -task classification, are discussed.

This work is very similar to Grimes et al.'s [31] work regarding task and in the attempt to classify across tasks. The two crucial differences are the physiological measure and the choice to build models for individuals. They investigated the feasibility of using EEG, while this work primarily investigated the feasibility of using pupil dilation. Their models were able to discriminate between four difficulty levels of the letter n-back task with 88% accuracy for individuals. Such specificity could neither be reached by individual models nor population models in this work. They report 77.05% and 80.4% when training the models on the letter task and testing it with two variants (images and spatial locations). These accuracies were achieved when discriminating between low (0-back) and high load (3-back). The population models in the present study reached up to 90.1% and 79.4% for these variants for $1v2$. These values cannot be compared directly but give an idea of what is possible for individual models, and thus, show that this work's cross-classification perform well even with a population model.

A very recent publication by Appel et al. [3] used a previous study's data of an n-back letter variant during which a remote eye tracker recorded pupil dilation.

Their goal as well was cross-user classification, yet, their approach was quite different from this work. While the population models of section 5 simply trained one model with all the data they built models for individuals, each model was given a similarity score to a new user. To make a prediction the most similar models were used for a weighted voting classification. As features they used the pupil's median, the *Index of Cognitive Activity* (ICA) [45] and blinks. Because of a limited amount of data, they chose small windows (one to five seconds) without overlap (step size = window size). They reach 71.5% for $1v2$ and a window size of five seconds. The best result in this work for $WIN5$ and $1v2$ was 82.6% - a difference of more than 10%. Their machine learning approach was rather different which might explain a better performance. In addition to that, in this work, multiple features (27 in total) derived from the pupil signal were used in contrast to three features in their study. Interestingly, they had poor performance when discriminating between 0-back and 1-back, comparable to the present issue for $2v3$ just at the other end of the difficulty spectrum.

Gevins et al. [28] used a spatial and letter n-back variant in their study to build individual and population models using EEG. One of another difference, besides the measure, was that they had four sessions for each participant, two practice days, one testing day and a retest day one month after testing. Only the testing and retest data was used for building models. Using retest additionally allowed them to have a cross-session classification. Their participants, as a result of the long practice, where supposed to be able to perform both tasks very well which rules out effects as occurred in the present work where 3-back most likely needed more practice time to develop a decent approach. They as well used 3-back and were able to discriminate between it and 2-back with an average accuracy of 80% (individual models). Also, they applied a window approach where the smallest window was represented by one trial. The latter is consisting of 200ms stimulus presentation a blank screen for 4.3 seconds, thus, 4.5 seconds long. In most cases multiple trials were used as one window, without overlap, e.g. the 80% accuracy was reached with nine of these trials (40.5-second windows). Using 27-second windows, they were able to discriminate between 1-back and 3-back with an accuracy of 94% when training with data of one variant and testing it with the other. It is assumed that it is the result of individual models but is not explained clearly in their work. They report 83% accuracy (13.5s windows) for cross-user classification when both task variants are treated as one data set. In the present work for $WIN10$, classification reached 85% for $1v3$ which indicates that both EEG and pupil dilation measures can be used for working memory load prediction. While they only had eight participants, they had much data for each (6-8 hours per session). Still, it could be argued that having eight participants is not enough to generalize the results.

It appears that often the results are not reported with full transparency from which the present approach could have benefited or so it could be compared better. For instance, classification results other than comparing $0v3$ in [31] are not reported for cross-task classification. This would have been useful for contrasting the present result. Furthermore, it is not mentioned how or if they account for between task variants' differences. If task variants do not produce the same physiological response, classification most likely is affected negatively if the data processing does not account for the differences.

To the extent to which the related work is comparable, it can be stated that the present work reaches similar promising results indicating the feasibility of using pupil dilation for cognitive state assessment.

7.7 Limitations

At the end of this thesis, the conducted research can be contrasted with the main research goal of classifying cognitive state using psychophysiological measures and which limitations are present in this work.
The first study was titled data gathering study as it was intended to collect ground truth data which can be used for classification. It was assumed that the difficulty levels represent low, medium and high working memory load. The task itself was new for participants. The practice phase allowed them to get familiar with it. Still, as 3-back showed, finding a suitable strategy takes time and most likely added load. Therefore, the ground truth data gathered might have been influenced by factors creating noise in the representation of the working memory state. Future experiments could improve this by having a much longer practice phase so the learning effect within a task difficulty can be reduced as done in [28] or [78]. Alternatively, the first trial runs of a difficulty level might be removed, for instance, Grimes et al. [31] removed the two first blocks which would as well require a longer total length of a condition to have enough data.

It was mentioned at several occasions that some scaling approaches or other data processing steps cannot be transferred directly to a real-time system. Thus, while this work uses some methodology (e.g. sliding windows) which are real-time suitable, the promising classification results will, for now, only be useful in off-line scenarios where all data to classify is available. The research objective could have been defined more precisely regarding that matter. The time invested in trying to satisfy both real-time and non-real-time approaches could have been used to focus on one of both aspects.

Using the pupil is very promising. However, the data gathering study was a very controlled experiment, and to which extent the findings can be reproduced in a less controlled and more general setting remains an open question. As proposed by Duchowski et al. [22] the illuminance measured in lux should be reported in an eye-tracking experiment to be able to reproduce the results which was not done in the present work. If it is desired to go out of the laboratory the light needs to be accounted for.

Furthermore, the baseline measure used (0-back condition) varies too much and does not always work as intended. Thus, a more traditional baseline at the beginning of each difficulty condition might have been more suitable. The time frame (20 seconds) for each 0-back phase in each trial run might as well not be suitable for a baseline.

Even though EDA and BVP measures were not significantly different in each difficulty condition, they could have been used for classification, to investigate, if anything useful can be harvested from the signals. Further, at least blinks could have been extracted from the pupil signal instead of averaging out missing values using smoothing. Further, performance measures could have been used to try and discriminate between 2-back and 3-back better.

8 CONCLUSION

Motivated by the promises of intelligent adaptive systems to improve user performance, to support the users' goals and reduce workload, this work investigated the use of unobtrusive physiological measures to classify the user state. The latter is one aspect of a general adaptive system. This work attempted to contribute to the realization of them by investigating the assessment of user state, in particular, the cognitive state. The classification of the cognitive state is not only beneficial for adaptive systems, but also for evaluation purposes. A reliable measure of the cognitive state could be used for evaluation studies instead of subjective post hoc measures like the NASA-TLX. The findings presented in this work might be used to develop an alternative. In the following, some directions for future work are presented.

To improve the present approach data processing could be extended to account for confounding factors such as light or display luminosity. Research isolating this challenge exists, e.g. [62]. Their findings could be used to create a more sophisticated classifier. Doing so would open up possibilities to investigate classification in the wild. Classification accuracy was high when the data was scaled with the z-score for individuals and each variant. It can be seen as an off-line approach. Therefore, the studies' data could be used to investigate online classification to get closer to a real-time system. The purpose of scaling is to account for differences across users and tasks. Future work should investigate alternative approaches, e.g. as attempted by Appel et al. [3]. The current model is not able to discriminate between a medium load (2-back) and a high load (3-back). Therefore, the specificity could be improved. The 2-back condition might be seen as a regular and good level of load. A safety-critical system where overload is not desired needs a detection of the point in time when the user state is shifting from a regular level to a too high level in order to be able to prevent overload. Future work can investigate if this is possible using only pupil diameter, if additional information from other physiological sensors is necessary or if more than physiological data is required(e.g. behavioural data) as suggested by [16].

The working memory task was used to elicit different levels of cognitive state which were assumed to represent low to high load. It is arguable if the n-back conditions represent low and elevated working memory load in general. However, assuming it does, the experimental setting in which the data was collected could be improved so that cleaner data is gathered. The term cleaner refers to data that is not affected by learning or fatigue effects. Also, the data of a condition should represent the effort necessary for it. One way to achieve this would be to let participants practice the difficulty conditions as long as they want. In addition, they could be told to develop a strategy that they think will work best for the condition. This might avoid participants switching strategies within a difficulty level. It would be interesting to see if another pupil pattern can be elicited during the 3-back condition.

The model could be validated by comparing it with existing workload measures. For instance, in an arbitrary experiment with two conditions, the output of the classifier could be compared with the NASA-TLX ratings of each condition. Assuming condition A results in a low TLX rating and the other (B) in a high rating, it would be expected that the model predominantly outputs low load for A and elevated load for B. If this is true then it can be assumed that the model can be used as an alternative to the NASA-TLX rating. The benefit of the population model would be a fine-grained workload estimation over the whole duration of the condition instead of one single measure at the end. The recent description sketches the idea of how future work could start to evaluate the population model so it can be used as an evaluation tool.

For now, a truly adaptive system which reliably detects the cognitive state cannot be built with the results of this work. Still, if the approach is made real-time suitable, it could be used for adaptation studies similar to Rajan et al. [64]. The latter compared performance measures of a condition in which their cognitive state classifier mediated notifications with a condition where they sent notifications randomly. This works model could be used for similar approaches.
This work focused on the working memory state; however, the user state can be much more than that. For instance, emotional states such as frustration and boredom can be used to describe the user state. Also, attention and engagement describe user state. These other states might as well be classified using physiological data. Therefore, to advance to an adaptive system which can react to the user state in general more facets of it have to be detected than just working memory state.

The cross-user and cross-task classification presented in this work is basic research and has to be improved and validated more before an intelligent system can be built that is able to monitor or react to the cognitive state of users in real-world systems. This work did show the feasibility of using pupil diameter for classification of working memory tasks which is a first and necessary step towards the building of such systems.

REFERENCES

[1] James L. Alty. 2003. *Handbook of Cognitive Task Design.* CRC Press, Chapter Cogntive Workload and Adaptive Systems, 129 – 145.

[2] Pavlo Antonenko, Fred Paas, Roland Grabner, and Tamara van Gog. 2010. Using Electroencephalography to Measure Cognitive Load. *Educational Psychology Review* 22, 4 (Dec. 2010), 425–438. https://doi.org/10.1007/s10648-010-9130-y

[3] Tobias Appel, Christian Scharinger, Peter Gerjets, and Enkelejda Kasneci. 2018. Cross-subject Workload Classification Using Pupil-related Measures. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA '18).* ACM, New York, NY, USA, 4:1–4:8. https://doi.org/10.1145/3204493.3204531

[4] Hasan Ayaz, Banu Onaral, Kurtulus Izzetoglu, Patricia A. Shewokis, Ryan McKendrick, and Raja Parasuraman. 2013. Continuous monitoring of brain dynamics with functional near infrared spectroscopy as a tool for neuroergonomic research: empirical examples and a technological development. *Frontiers in Human Neuroscience* 7 (2013). https://doi.org/10.3389/fnhum.2013.00871

[5] Alan Baddeley. 1992. Working Memory: The Interface between Memory and Cognition. *Journal of Cognitive Neuroscience* 4, 3 (July 1992), 281–288. https://doi.org/10.1162/jocn.1992.4.3.281

[6] Carryl L. Baldwin and B.N. Penaranda. 2012. Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification. *NeuroImage* 59, 1 (Jan. 2012), 48–56. https://doi.org/10.1016/j.neuroimage.2011.07.047

[7] Victor Manuel García Barrios, Christian Gütl, Alexandra M. Preis, Keith Andrews, Maja Pivec, Felix Mödritscher, and Christian Trummer. 2004. AdELE: A framework for adaptive e-learning through eye tracking. *Proceedings of IKNOW* (2004), 609–616. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.60.1863&rep=rep1&type=pdf

[8] Jackson Beatty and Brennis Lucero-Wagoner. 2000. The Pupillary System. In *Handbook of Psychophysiology* (second ed.), John T. Cacioppo, Louis G. Tassinary, and Gary G. Berntson (Eds.). Cambridge University Press, Cambridge, UK ; New York, NY, USA, Chapter 6. http://www.nrc-iol.org/cores/mialab/fijc/files/2003/090203_Pupillary_System_.pdf

[9] Gary G. Berntson, John T. Cacioppo, and Jos A. Bosch. 2017. From Homeostasis to Allodynamic Regulation. In *Handbook of Psychophysiology* (4 ed.), John T. Cacioppo, Louis G. Tassinary, and Gary G. Berntson (Eds.). Cambridge University Press, Cambridge, 401–426. http://ebooks.cambridge.org/ref/id/CBO9781107415782A031 DOI: 10.1017/9781107415782.018.

[10] Gary G. Berntson, Karen S. Quigley, Greg J. Norman, and David L. Lozano. 2017. Cardiovascular Psychophysiology. In *Handbook of Psychophysiology* (4 ed.), John T. Cacioppo, Louis G. Tassinary, and Gary G. Berntson (Eds.). Cambridge University Press, Cambridge, 183–216. http://ebooks.cambridge.org/ref/id/CBO9781107415782A021 DOI: 10.1017/9781107415782.009.

[11] Gérard Biau and Erwan Scornet. 2016. A random forest guided tour. *TEST* 25, 2 (June 2016), 197–227. https://doi.org/10.1007/s11749-016-0481-7

[12] Anne-Laure Boulesteix, Silke Janitza, Jochen Kruppa, and Inke R. König. 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2, 6 (2012), 493–507. https://doi.org/10.1002/widm.1072

[13] Margaret M. Bradley, Laura Miccoli, Miguel A. Escrig, and Peter J. Lang. 2008. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology* 45, 4 (July 2008), 602–607. https://doi.org/10.1111/j.1469-8986.2008.00654.x

[14] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (Oct. 2001), 5–32. https://doi.org/10.1023/A:1010933404324

[15] John T. Cacioppo, Lousi G. Tassinary, and Gary G. Berntson. 2016. *Handbook of Psychophysiology* (4 ed.). Cambridge University Press. https://doi.org/10.1017/9781107415782

[16] Fang Chen, Jianlong Zhou, Yang Wang, Kun Yu, Syed Z. Arshad, Ahmad Khawaji, and Dan Conway. 2016. *Robust Multimodal Cognitive Load Measurement.* Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-31700-7

[17] Siyuan Chen, Julien Epps, and Fang Chen. 2013. Automatic and Continuous User Task Analysis via Eye Activity. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI '13).* ACM, New York, NY, USA, 57–66. https://doi.org/10.1145/2449396.2449406

[18] Siyuan Chen, Julien Epps, Natalie Ruiz, and Fang Chen. 2011. Eye Activity As a Measure of Human Mental Effort in HCI. In *Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI '11).* ACM, New York, NY, USA, 315–318. https://doi.org/10.1145/1943403.1943454

[19] Andrew R. A. Conway, Michael J. Kane, Michael F. Bunting, D. Zach Hambrick, Oliver Wilhelm, and Randall W. Engle. 2005. Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review* 12, 5 (Oct. 2005), 769–786. https://doi.org/10.3758/BF03196772

[20] Benjamin Cowley, Marco Filetti, Kristian Lukander, Jari Torniainen, Andreas Henelius, Lauri Ahonen, Oswald Barral, Ilkka Kosunen, Teppo Valtonen, Minna Huotilainen, Niklas Ravaja, and Giulio Jacucci. 2016. The Psychophysiology Primer: A Guide to Methods and a Broad Review with a Focus on Human–Computer Interaction. *Foundations and Trends® in Human–Computer Interaction* 9, 3-4 (Nov. 2016), 151–308. https://doi.org/10.1561/1100000065

[21] Michael E. Dawson, Anne M. Schell, and Diane L. Filion. 2017. The Electrodermal System. In *Handbook of Psychophysiology* (4 ed.), John T. Cacioppo, Louis G. Tassinary, and Gary G. Berntson (Eds.). Cambridge University Press, Cambridge, 217–243. http://ebooks.cambridge.org/ref/id/CBO9781107415782A022 DOI: 10.1017/9781107415782.010.

[22] Andrew T. Duchowski, Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos. 2018. The Index of Pupillary Activity: Measuring Cognitive Load Vis-à-vis Task Difficulty with Pupil Oscillation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18).* ACM, New York, NY, USA, 282:1–282:13. https://doi.org/10.1145/3173574.3173856

[23] Mai Elkomy, Yomna Abdelrahman, Markus Funk, Tilman Dingler, Albrecht Schmidt, and Slim Abdennadher. 2017. ABBAS: An Adaptive Bio-sensors Based Assistive System. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17).* ACM, New York, NY, USA, 2543–2550. https://doi.org/10.1145/3027063.3053179

[24] Stephen H. Fairclough. 2009. Fundamentals of physiological computing. *Interacting with Computers* 21, 1-2 (Jan. 2009), 133–145. https://doi.org/10.1016/j.intcom.2008.10.011

[25] Karen M. Feigh, Michael C. Dorneich, and Caroline C. Hayes. 2012. Toward a Characterization of Adaptive Systems: A Framework for Researchers and System Designers. *Human Factors* 54, 6 (Dec. 2012), 1008–1024. https://doi.org/10.1177/0018720812443983

[26] E. Ferreira, D. Ferreira, S. Kim, P. Siirtola, J. Röning, J. F. Forlizzi, and A. K. Dey. 2014. Assessing real-time cognitive load based on psycho-physiological measures for younger and older adults. In *IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB).* 39–48. https://doi.org/10.1109/CCMB.2014.7020692

[27] Thomas M. Gable, Andrew L. Kun, Bruce N. Walker, and Riley J. Winton. 2015. Comparing Heart Rate and Pupil Size As Objective Measures of Workload in the Driving Context: Initial Look. In *Adjunct Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '15).* ACM, New York, NY, USA, 20–25.

https://doi.org/10.1145/2809730.2809745

[28] Alan Gevins, Michael E. Smith, Harrison Leong, Linda McEvoy, Susan Whitfield, Robert Du, and Georgia Rush. 1998. Monitoring Working Memory Load during Computer-Based Tasks with EEG Pattern Recognition Methods , Monitoring Working Memory Load during Computer-Based Tasks with EEG Pattern Recognition Methods. *Human Factors* 40, 1 (March 1998), 79–91. https://doi.org/10.1518/001872098779480578

[29] Mariel Grassmann, Elke Vlemincx, Andreas von Leupoldt, Mittelst&#xe4, Justin M. Dt, and Omer Van den Bergh. 2016. Respiratory Changes in Response to Cognitive Load: A Systematic Review. *Neural Plasticity* (2016). https://www.hindawi.com/journals/np/2016/8146809/abs/ DOI: 10.1155/2016/8146809.

[30] Gabriele Gratton and Monica Fabiani. 2017. Biosignal Processing in Psychophysiology: Principles and Current Developments. In *Handbook of Psychophysiology* (4 ed.), John T. Cacioppo, Louis G. Tassinary, and Gary G. Berntson (Eds.). Cambridge University Press, Cambridge, 628–661. http://ebooks.cambridge.org/ref/id/CBO9781107415782A043 DOI: 10.1017/9781107415782.029.

[31] David Grimes, Desney S. Tan, Scott E. Hudson, Pradeep Shenoy, and Rajesh P.N. Rao. 2008. Feasibility and Pragmatics of Classifying Working Memory Load with an Electroencephalograph. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 835–844. https://doi.org/10.1145/1357054.1357187

[32] Eija Haapalainen, SeungJun Kim, Jodi F. Forlizzi, and Anind K. Dey. 2010. Psycho-physiological Measures for Assessing Cognitive Load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp '10)*. ACM, New York, NY, USA, 301–310. https://doi.org/10.1145/1864349.1864395

[33] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology* 52 (Jan. 1988), 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[34] Nina Hollender, Cristian Hofmann, Michael Deneke, and Bernhard Schmitz. 2010. Integrating cognitive load theory and concepts of human–computer interaction. *Computers in Human Behavior* 26, 6 (Nov. 2010), 1278–1288. https://doi.org/10.1016/j.chb.2010.05.031

[35] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost van de Weijer. 2011. *Eye Tracking: A comprehensive guide to methods and measures*. OUP Oxford. http://www.ebook.de/de/product/21364941/kenneth_holmqvist_marcus_nystr_ouml_m_richard_andersson_richard_dewhurst_halszka_jarodzka_eye_tracking_a_comprehensive_guide_to_methods_and_measures.html Google-Books-ID: 5rIDPV1EoLUC.

[36] Shamsi T. Iqbal, Piotr D. Adamczyk, Xianjun Sam Zheng, and Brian P. Bailey. 2005. Towards an Index of Opportunity: Understanding Changes in Mental Workload During Task Execution. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05)*. ACM, New York, NY, USA, 311–320. https://doi.org/10.1145/1054972.1055016

[37] Arthur Kramer. 1991. *Multiple-task Performance*. Taylor & Francis, Chapter Physiological Metrics of Mental Workload: A Review of Recent Progress, 279 – 328. https://pdfs.semanticscholar.org/4438/5dc58bf688e65af48c1a8b6fa55b80268c58.pdf

[38] Peter J. Lang, Mark K. Greenwald, Margaret M. Bradley, and Alfons O. Hamm. 1993. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology* 30, 3 (May 1993), 261–273. https://doi.org/10.1111/j.1469-8986.1993.tb03352.x

[39] Can Liu, Olivier Chapuis, Michel Beaudouin-Lafon, Eric Lecolinet, and Wendy E. Mackay. 2014. Effects of display size and navigation type on a classification task. In *In Proceedings of the*

32nd Annual ACM Conference on Human Factors in Computing Systems. ACM Press, 4147–4156. https://doi.org/10.1145/2556288.2557020

[40] Tyler S. Lorig. 2017. The Respiratory System. In *Handbook of Psychophysiology* (4 ed.), John T. Cacioppo, Louis G. Tassinary, and Gary G. Berntson (Eds.). Cambridge University Press, Cambridge, 244–257. http://ebooks.cambridge.org/ref/id/CBO9781107415782A023 DOI: 10.1017/9781107415782.011.

[41] Gilles Louppe. 2014. Understanding Random Forests: From Theory to Practice. *arXiv:1407.7502 [stat]* (July 2014). http://arxiv.org/abs/1407.7502 arXiv: 1407.7502.

[42] Angela Mahr, Michael Feld, Mohammad Mehdi Moniri, and Rafael Math. 2012. The contre (continuous tracking and reaction) task: A flexible approach for assessing driver cognitive workload with high sensitivity. *Adjunct Proceedings of the 4th AutomotiveUI, Portsmouth. ACM* (2012). http://www.auto-ui.org/12/adjunct-proceedings/w2-06-mahr.pdf

[43] Marek Malik. 1996. Heart Rate Variability. *Annals of Noninvasive Electrocardiology* 1, 2 (April 1996), 151–181. https://doi.org/10.1111/j.1542-474X.1996.tb00275.x

[44] Robert Malmivuo, Jaakko; Plonsey. 1995. *Bioelectromagnetism: principles and applications of bioelectric and biomagnetic fields*. Oxford University Press. http://www.bem.fi/book/

[45] S.P. Marshall. 2002. The Index of Cognitive Activity: measuring cognitive workload. In *Proceedings of the 2002 IEEE 7th Conference on Human Factors and Power Plants*. IEEE, 7–5–7–9. https://doi.org/10.1109/HFPP.2002.1042860

[46] Stewart Martin. 2014. Measuring cognitive load and cognition: metrics for technology-enhanced learning. *Educational Research and Evaluation* 20, 7-8 (Nov. 2014), 592–621. https://doi.org/10.1080/13803611.2014.997140

[47] Sebastiaan Mathôt, Daniel Schreij, and Jan Theeuwes. 2012. OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods* 44, 2 (2012), 314–324.

[48] George A. Miller. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* 63, 2 (1956), 81–97. https://doi.org/10.1037/h0043158

[49] Ron Milo, Paul Jorgensen, Uri Moran, Griffin Weber, and Michael Springer. 2010. BioNumbers—the database of key numbers in molecular and cell biology. *Nucleic Acids Research* 38, Database issue (Jan. 2010), D750–D753. https://doi.org/10.1093/nar/gkp889

[50] Neville Moray. 1979. Models and Measures of Mental Workload. In *Mental Workload*. Springer, Boston, MA, 13–21. https://link.springer.com/chapter/10.1007/978-1-4757-0884-4_2 DOI: 10.1007/978-1-4757-0884-4_2.

[51] Shane T. Mueller and Brian J. Piper. 2014. The Psychology Experiment Building Language (PEBL) and PEBL Test Battery. *Journal of Neuroscience Methods* 222, Supplement C (Jan. 2014), 250–259. https://doi.org/10.1016/j.jneumeth.2013.10.024

[52] Lennart Nacke. 2009. *Affective Ludology : Scientific Measurement of User Experience in Interactive Entertainment*. Ph.D. Dissertation. Blekinge Institute of Technology, School of Computing. http://www.diva-portal.org/smash/record.jsf?pid=diva2:835627

[53] Lennart E. Nacke. 2013. An Introduction to Physiological Player Metrics for Evaluating Games. In *Game Analytics*, Magy Seif El-Nasr, Anders Drachen, and Alessandro Canossa (Eds.). Springer London, 585–619. http://link.springer.com/chapter/10.1007/978-1-4471-4769-5_26 DOI: 10.1007/978-1-4471-4769-5_26.

[54] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo. 2015. Recognizing Emotions Induced by Affective Sounds through Heart Rate Variability. *IEEE Transactions on Affective Computing* 6, 4 (Oct. 2015), 385–394. https://doi.org/10.1109/TAFFC.2015.2432810

[55] Nargess Nourbakhsh, Yang Wang, and Fang Chen. 2013. GSR and blink features for cognitive load classification. In *IFIP Conference on Human-Computer Interaction*. Springer, 159–166. http://link.springer.com/chapter/10.1007/978-3-642-40483-2_11

[56] Nargess Nourbakhsh, Yang Wang, Fang Chen, and Rafael A. Calvo. 2012. Using Galvanic Skin Response for Cognitive Load Measurement in Arithmetic and Reading Tasks. In *Proceedings of the 24th Australian Computer-Human Interaction Conference (OzCHI '12)*. ACM, New York, NY, USA, 420–423. https://doi.org/10.1145/2414536.2414602

[57] Sharon Oviatt. 2006. Human-centered design meets cognitive load theory: designing interfaces that help people think. In *Proceedings of the 14th ACM International Conference on Multimedia*. ACM, 871–880. http://dl.acm.org/citation.cfm?id=1180831

[58] Fred Paas, Juhani E. Tuovinen, Huib Tabbers, and Pascal W. M. Van Gerven. 2003. Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist* 38, 1 (March 2003), 63–71. https://doi.org/10.1207/S15326985EP3801_8

[59] Timo Partala, Maria Jokiniemi, and Veikko Surakka. 2000. Pupillary responses to emotionally provocative stimuli. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*. ACM, 123–129. http://dl.acm.org/citation.cfm?id=355042

[60] Ian Peate and Muralitharan Nair. 2011. *Fundamentals of Anatomy and Physiology for Student Nurses*. John Wiley & Sons. Google-Books-ID: 8w6nmuTp3MEC.

[61] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[62] Bastian Pfleging, Drea K. Fekety, Albrecht Schmidt, and Andrew L. Kun. 2016. [SKIMMED] A Model Relating Pupil Diameter to Mental Workload and Lighting Conditions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5776–5788. https://doi.org/10.1145/2858036.2858117

[63] Felix Putze and Tanja Schultz. 2014. Adaptive cognitive technical systems. *Journal of Neuroscience Methods* 234 (Aug. 2014), 108–115. https://doi.org/10.1016/j.jneumeth.2014.06.029

[64] Rahul Rajan, Ted Selker, and Ian Lane. 2016. Task Load Estimation and Mediation Using Psychophysiological Measures. In *Proceedings of the 21st International Conference on Intelligent User Interfaces (IUI '16)*. ACM, New York, NY, USA, 48–59. https://doi.org/10.1145/2856767.2856769

[65] Sebastian Raschka and Vahid Mirjalili. 2017. *Python Machine Learning*. Packt Publishing Ltd.

[66] W. Ray and H. Cole. 1985. EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes. *Science* 228, 4700 (May 1985), 750–752. https://doi.org/10.1126/science.3992243

[67] Dennis W Rowe, John Sibert, and Don Irwin. 1998. Heart rate variability: Indicator of user state as an aid to human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press/Addison-Wesley Publishing Co., 480–487.

[68] Abigail Sellen, Yvonne Rogers, Richard Harper, and Tom Rodden. 2009. Reflecting Human Values in the Digital Age. *Commun. ACM* 52, 3 (March 2009), 58–66. https://doi.org/10.1145/1467247.1467265

[69] Yu Shi, Natalie Ruiz, Ronnie Taib, Eric Choi, and Fang Chen. 2007. Galvanic skin response (GSR) as an index of cognitive load. In *CHI'07 Extended Abstracts on Human Factors In Computing Systems*. ACM, 2651–2656.

[70] Erin T. Solovey, Marin Zec, Enrique Abdon Garcia Perez, Bryan Reimer, and Bruce Mehler. 2014. Classifying Driver Workload Using Physiological and Driving Performance Data: Two Field Studies. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing*

Systems (CHI '14). ACM, New York, NY, USA, 4057–4066. https://doi.org/10.1145/2556288.2557068

[71] John Sweller, Jeroen JG Van Merrienboer, and Fred GWC Paas. 1998. Cognitive architecture and instructional design. *Educational Psychology Review* 10, 3 (1998), 251–296. http://link.springer.com/article/10.1023/A:1022193728205

[72] Maria Teresa Valderas, Juan Bolea, Pablo Laguna, Montserrat Vallverdú, and Raquel Bailón. 2015. Human emotion recognition using heart rate variability analysis with spectral bands based on respiration. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 6134–6137. http://ieeexplore.ieee.org/abstract/document/7319792/

[73] Philipp von Bauer. 2017. Master Seminar: Towards real-time cognitive state assessment using psychophysiological measures. (Nov. 2017).

[74] Philipp von Bauer. 2018. Master Project Report: Towards Classifying Cognitive State. (April 2018).

[75] Carina Walter, Stephanie Schmidt, Wolfgang Rosenstiel, Peter Gerjets, and Martin Bogdan. 2013. Using Cross-Task Classification for Classifying Workload Levels in Complex Learning Tasks. In *Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 876–881. https://doi.org/10.1109/ACII.2013.164

[76] Weihong Wang, Zhidong Li, Yang Wang, and Fang Chen. 2013. Indexing Cognitive Workload Based on Pupillary Response Under Luminance and Emotional Changes. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI '13)*. ACM, New York, NY, USA, 247–256. https://doi.org/10.1145/2449396.2449428

[77] E. J. Williams. 1949. Experimental Designs Balanced for the Estimation of Residual Effects of Treatments. *Australian Journal of Scientific Research A Physical Sciences* 2 (June 1949), 149. https://doi.org/10.1071/PH490149

[78] Glenn F. Wilson and Christopher A. Russell. 2003. Real-Time Assessment of Mental Workload Using Psychophysiological Measures and Artificial Neural Networks. *Human Factors* 45, 4 (Dec. 2003), 635–644. https://doi.org/10.1518/hfes.45.4.635.27088

[79] Johannes Zagermann, Ulrike Pfeil, Daniel Fink, Philipp von Bauer, and Harald Reiterer. 2017. Memory in Motion: The Influence of Gesture- and Touch-Based Input Modalities on Spatial Memory. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1899–1910. https://doi.org/10.1145/3025453.3026001

[80] Jianlong Zhou, Syed Z. Arshad, Simon Luo, Kun Yu, Shlomo Berkovsky, and Fang Chen. 2017. Indexing Cognitive Load Using Blood Volume Pulse Features. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2269–2275. https://doi.org/10.1145/3027063.3053140

[81] Jianlong Zhou, Ju Young Jung, and Fang Chen. 2015. Dynamic Workload Adjustments in Human-Machine Systems Based on GSR Features. In *Human-Computer Interaction – INTERACT 2015*. Springer, Cham, 550–558. https://doi.org/10.1007/978-3-319-22701-6_40

[82] Jianlong Zhou, Jinjun Sun, Fang Chen, Yang Wang, Ronnie Taib, Ahmad Khawaji, and Zhidong Li. 2015. Measurable Decision Making with GSR and Pupillary Analysis for Intelligent User Interface. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 6 (Jan. 2015), 33:1–33:23. https://doi.org/10.1145/2687924

A PERFORMANCE, TLX STUDY 1

A.1 NASA TLX Descriptives

	TLX Raw			TLX Raw without physical demand		
	1-back	2-back	3-back	1-back	2-back	3-back
Mean	32.604	47.326	57.014	36.625	53.333	64.708
Std. Error of Mean	2.724	2.576	2.957	3.061	2.939	3.153
Median	31.667	50.417	58.333	34.500	57.500	65.500
Std. Deviation	13.343	12.619	14.488	14.995	14.397	15.448
Minimum	10.000	15.833	18.333	8.000	19.000	22.000
Maximum	63.333	65.000	91.667	66.000	76.000	91.000

Table A1. Descriptive statistics of the NASA-TLX ratings.

A.2 Performance Descriptive

	Response Accuracy			Response Time		
	1-back	2-back	3-back	1-back	2-back	3-back
Mean	0.968	0.896	0.770	583.554	739.276	836.635
Std. Error of Mean	0.005	0.012	0.015	16.717	23.290	23.540
Median	0.976	0.917	0.770	559.335	731.388	843.867
Std. Deviation	0.024	0.061	0.072	81.896	114.096	115.322
Minimum	0.887	0.742	0.621	470.790	563.677	621.387
Maximum	0.992	0.976	0.919	754.323	1016.266	1105.113

Table A2. Descriptive statistics of the performance scores (accuracy and response time).

A.3 NASA TLX Friedman Test and Post Hoc

Factor	Chi-Squared	df	p	Kendall's W
RAW	28.667	2	< .001	0.566

Table A3. Friedman Test

			W	p	Rank-Biserial Correlation
n1_raw	-	n2_raw	26.000	0.001	-0.827
	-	n3_raw	7.000	< .001	-0.953
n2_raw	-	n3_raw	24.000	< .001	-0.840

Table A4. Wilcoxon

A.4 Performance Friedman Test and Post Hoc

Factor	Chi-Squared	df	p	Kendall's W
acc	42.250	2	< .001	0.496
time	42.750	2	< .001	0.719

Table A5. Friedman Test

			W	p	Rank-Biserial Correlation
response_accuracy_n1	-	response_accuracy_n2	291.500	< .001	0.943
	-	response_accuracy_n3	300.000	< .001	1.000
response_accuracy_n2	-	response_accuracy_n3	297.500	< .001	0.983
average_response_time_n1	-	average_response_time_n2	0.000	< .001	-1.000
	-	average_response_time_n3	0.000	< .001	-1.000
average_response_time_n2	-	average_response_time_n3	21.000	< .001	-0.860

Table A6. Wilcoxon

B PUPIL ANALYSIS STUDY 1

B.1 Quality Inspection

	Raw		Smoothed		Smoothed (Rolling)	Smoothed (Hanning)	Remarks
Participant	L	R	L	R	L+R	L+R	
1	-	-	o	o	oV	oV	
2	-	-	o	o	+	+	n1_t1 x
3	x	-	-	-	-	o	n3 x, only R
4	-	-	o	o	+	+	n1_t3 -?
5	-	-	o	o	+	o	
6	-	-	-	-	o	o	
7	x?	x?	-x?	-x?	oV	+	
8	-x?	-	-	o	o	+	n2/n3 -
9	-x?	-x?	-x?	-x?	oV	o	
10	-	-	-x?	oV	o	+	only R
11	-x?	-	-x?	oV	-V	o(R)	only R
12	-o	-o	oV	oV	oV	+	
13	-	-	o	o	+	+	only L
14	-	x?	o	-	+(L)	o(L)	n3 -V
15	-	-V	o	o	+	+	
16	-	-V	-o	-o	o	o	
17	-	-	+	+	+	+	
19	o	o	+	+	+	o	
20	-	-x?	o	-o	+	+	only L?
21	-o	-0	o	o	+	o	n3_t4 pupil/bvp raise?
22	-	-x?	o	-o	+	+	(3,3,3) missing
23	-	-	o-	o	+	+	gaps in HR
24	-o	-o	oV	oV	+	+	lenovo update n1_t4
31	-	-	+	+	+	+	
33	-	-	+	+	+	+	
35	-	-	+	+	+	+	
36	-	-	+	+	+	+	
39	-	-	o	o	+	+	
43	-	-	+	+	+	+	
49	-	-	+	+	+	+	

Table A7. Checking Participant Physiological Data

x : kick it, - : noisy, o : okay, + : good, V: high variance, ? : take second look, L : left eye, R : right eye
Raw: using the raw pupil signal
Smoothed: rolling window using the median
Smoothed (Rolling): rolling window using the median, outliers removed in advance
Smoothed (Hanning): using a Hanning window, convolving the signal, outliers removed in advance

Participant	n1<n2<n3	<0.08	< 0.1
1	+	o	n3
2	o	o	o
3	+	+	+
4	+	o	o
5	+	o	o
6	+	o	o
7	-	-	-
8	o	o	o
9	-	-	-
10	+	o	o
11	o	-	-
12	+	o	o
13	+	o	o
14	n3	n3	n3
15	+	+	+
16	o	o	n2
17	+	o	o
19	+	o	o
20	+	o	o
21	o	o	o
22	+	+	+
23	o	o	o
24	o	o	o
31	+	+	o
33	o	o	o
35	+	+	n2
36	o	n2	n2
39	+	+	+
43	-	-	-
49	o	o	o

Table A8. Checking Participant Pupil Pattern Mean

+: n1<n2<n3 is true
o: n1<n2 and n1<n3 is true
n2: only n1<n2 is true
n3: only n1<n3 is true
-: none of the above

B.2 ANOVA and Friedman Test

Repeated Measures ANOVA

	Sum of Squares	df	Mean Square	F	p	η^2	${\eta_p}^2$
MEAN	1.048	2	0.524	85.710	< .001	0.788	0.788
MAX	0.186	2	0.093	29.507	< .001	0.562	0.562
Q25	1.458	2	0.729	81.830	< .001	0.781	0.781

Friedman Test

	df	Chi-Squared	p	Kendall's W
MIN	2	27.750	< .001	0.827
P2P	2	16.750	< .001	0.574
MEDIAN	2	36.750	< .001	0.926
Q75	2	38.083	< .001	0.943
IQR	2	25.083	< .001	0.607
STD	2	28.083	< .001	0.683

Table A9. without n0 calibration

Repeated Measures ANOVA

	Sum of Squares	df	Mean Square	F	p	η^2	${\eta_p}^2$
MEAN	0.075	2	0.038	54.669	< .001		
MEDIAN	0.072	2	0.036	56.971	< .001		
P2P	0.042	2	0.021	5.654	0.006		
Q25	0.081	2	0.041	67.571	< .001		
STD	0.002	2	9.160e-4	5.720	0.006		

Friedman Test

	df	Chi-Squared	p	Kendall's W
MIN	2	25.333	< .001	0.571
IQR	2	4.333	0.115	0.573
MAX	2	12.333	0.002	0.582
Q75	2	30.583	< .001	0.599

Table A10. 0-back calibration

B.3 Assumption Checks

Without 0-back calibration

	Mauchly's W	p	Greenhouse-Geisser ϵ	Huynh-Feldt ϵ
MEAN	0.826	0.122	0.852	0.913
MEDIAN	0.884	0.258	0.896	0.967
MAX	0.825	0.121	0.851	0.912
MIN	0.473	$< .001$	0.655	0.678
P2P	0.508	$< .001$	0.670	0.696
Q25	0.762	0.050	0.808	0.860
Q75	0.919	0.393	0.925	1.000
IQR	0.945	0.537	0.948	1.000
STD	0.769	0.056	0.812	0.865

Using 0-back calibration

	Mauchly's W	p	Greenhouse-Geisser ϵ	Huynh-Feldt ϵ
MEAN	0.979	0.793	0.980	1.000
MEDIAN	0.963	0.662	0.965	1.000
MAX	0.904	0.328	0.912	0.987
MIN	0.744	0.039	0.796	0.846
P2P	0.969	0.709	0.970	1.000
Q25	0.965	0.677	0.966	1.000
Q75	0.784	0.068	0.822	0.877
IQR	0.602	0.004	0.715	0.749
STD	0.926	0.431	0.931	1.000

Table A11. Mauchly's Test of Spericity

			W	p
n1_mean	-	n2_mean	0.944	0.197
	-	n3_mean	0.971	0.702
n2_mean	-	n3_mean	0.949	0.252
n1_median	-	n2_median	0.809	< .001
	-	n3_median	0.953	0.309
n2_median	-	n3_median	0.953	0.313
n1_max	-	n2_max	0.947	0.230
	-	n3_max	0.916	0.049
n2_max	-	n3_max	0.946	0.220
n1_min	-	n2_min	0.959	0.417
	-	n3_min	0.946	0.224
n2_min	-	n3_min	0.907	0.030
n1_P2P	-	n2_P2P	0.898	0.020
	-	n3_P2P	0.969	0.651
n2_P2P	-	n3_P2P	0.936	0.132
n1_quantile25	-	n2_quantile25	0.955	0.354
	-	n3_quantile25	0.972	0.714
n2_quantile25	-	n3_quantile25	0.966	0.574
n1_quantile75	-	n2_quantile75	0.899	0.021
	-	n3_quantile75	0.969	0.639
n2_quantile75	-	n3_quantile75	0.943	0.187
n1_IQR	-	n2_IQR	0.864	0.004
	-	n3_IQR	0.943	0.189
n2_IQR	-	n3_IQR	0.782	< .001
n1_std	-	n2_std	0.916	0.048
	-	n3_std	0.944	0.200
n2_std	-	n3_std	0.792	< .001

			W	p
n1_calib_mean	-	n2_calib_mean	0.957	0.382
	-	n3_calib_mean	0.985	0.965
n2_calib_mean	-	n3_calib_mean	0.941	0.173
n1_calib_median	-	n2_calib_median	0.975	0.784
	-	n3_calib_median	0.965	0.540
n2_calib_median	-	n3_calib_median	0.934	0.120
n1_calib_max	-	n2_calib_max	0.905	0.027
	-	n3_calib_max	0.977	0.825
n2_calib_max	-	n3_calib_max	0.937	0.137
n1_calib_min	-	n2_calib_min	0.976	0.804
	-	n3_calib_min	0.933	0.117
n2_calib_min	-	n3_calib_min	0.892	0.014
n1_calib_P2P	-	n2_calib_P2P	0.960	0.444
	-	n3_calib_P2P	0.964	0.520
n2_calib_P2P	-	n3_calib_P2P	0.952	0.296
n1_calib_quantile25	-	n2_calib_quantile25	0.978	0.857
	-	n3_calib_quantile25	0.956	0.357
n2_calib_quantile25	-	n3_calib_quantile25	0.954	0.327
n1_calib_quantile75	-	n2_calib_quantile75	0.908	0.032
	-	n3_calib_quantile75	0.957	0.375
n2_calib_quantile75	-	n3_calib_quantile75	0.934	0.119
n1_calib_IQR	-	n2_calib_IQR	0.917	0.051
	-	n3_calib_IQR	0.955	0.351
n2_calib_IQR	-	n3_calib_IQR	0.909	0.033
n1_calib_std	-	n2_calib_std	0.948	0.245
	-	n3_calib_std	0.956	0.365
n2_calib_std	-	n3_calib_std	0.925	0.074

Table A12. Test of Normality (Shapiro-Wilk): on the left without using 0-back calibration, on the right using 0-back calibration

B.4 Descriptive Statistics

	N	Mean	SD	SE
n1_mean	24	3.143	0.324	0.066
n2_mean	24	3.380	0.271	0.055
n3_mean	24	3.415	0.295	0.060
n1_median	24	3.151	0.327	0.067
n2_median	24	3.392	0.276	0.056
n3_median	24	3.421	0.296	0.060
n1_max	24	3.469	0.303	0.062
n2_max	24	3.549	0.270	0.055
n3_max	24	3.591	0.290	0.059
n1_min	24	2.746	0.365	0.075
n2_min	24	3.109	0.254	0.052
n3_min	24	3.054	0.368	0.075
n1_P2P	24	0.723	0.232	0.047
n2_P2P	24	0.440	0.164	0.033
n3_P2P	24	0.538	0.248	0.051
n1_quantile25	24	3.043	0.340	0.069
n2_quantile25	24	3.319	0.275	0.056
n3_quantile25	24	3.365	0.297	0.061
n1_quantile75	24	3.248	0.321	0.066
n2_quantile75	24	3.447	0.277	0.057
n3_quantile75	24	3.475	0.296	0.060
n1_IQR	24	0.205	0.086	0.017
n2_IQR	24	0.128	0.083	0.017
n3_IQR	24	0.110	0.052	0.011
n1_std	24	0.149	0.059	0.012
n2_std	24	0.090	0.047	0.010
n3_std	24	0.087	0.034	0.007

	N	Mean	SD	SE
n1_calib_mean	24	-0.007	0.023	0.005
n2_calib_mean	24	0.060	0.040	0.008
n3_calib_mean	24	0.063	0.036	0.007
n1_calib_median	24	-0.004	0.019	0.004
n2_calib_median	24	0.062	0.040	0.008
n3_calib_median	24	0.065	0.035	0.007
n1_calib_max	24	0.093	0.041	0.008
n2_calib_max	24	0.131	0.064	0.013
n3_calib_max	24	0.129	0.054	0.011
n1_calib_min	24	-0.129	0.064	0.013
n2_calib_min	24	-0.032	0.044	0.009
n3_calib_min	24	-0.054	0.082	0.017
n1_calib_P2P	24	0.222	0.074	0.015
n2_calib_P2P	24	0.164	0.072	0.015
n3_calib_P2P	24	0.183	0.075	0.015
n1_calib_quantile25	24	-0.033	0.025	0.005
n2_calib_quantile25	24	0.034	0.034	0.007
n3_calib_quantile25	24	0.042	0.033	0.007
n1_calib_quantile75	24	0.023	0.017	0.004
n2_calib_quantile75	24	0.087	0.052	0.011
n3_calib_quantile75	24	0.085	0.039	0.008
n1_calib_IQR	24	0.056	0.018	0.004
n2_calib_IQR	24	0.054	0.036	0.007
n3_calib_IQR	24	0.043	0.018	0.004
n1_calib_std	24	0.044	0.015	0.003
n2_calib_std	24	0.035	0.019	0.004
n3_calib_std	24	0.032	0.011	0.002

Table A13. Descriptive Statistics: on the left without using 0-back calibration, on the right using 0-back calibration

B.5 Post Hoc Tests

			T-test			
			t	df	p	Cohen's d
n1_calib_mean	-	n2_calib_mean	-8.706	23	< .001	-1.777
	-	n3_calib_mean	-9.938	23	< .001	-2.028
n2_calib_mean	-	n3_calib_mean	-0.301	23	0.766	-0.062
n1_calib_median	-	n2_calib_median	-8.405	23	< .001	-1.716
	-	n3_calib_median	-10.433	23	< .001	-2.130
n2_calib_median	-	n3_calib_median	-0.437	23	0.666	-0.089
n1_calib_P2P	-	n2_calib_P2P	3.636	23	0.001	0.742
	-	n3_calib_P2P	2.121	23	0.045	0.433
n2_calib_P2P	-	n3_calib_P2P	-1.060	23	0.300	-0.216
n1_calib_quantile25	-	n2_calib_quantile25	-10.445	23	< .001	-2.132
	-	n3_calib_quantile25	-9.927	23	< .001	-2.026
n2_calib_quantile25	-	n3_calib_quantile25	-1.091	23	0.287	-0.223
n1_calib_std	-	n2_calib_std	2.339	23	0.028	0.477
	-	n3_calib_std	3.791	23	< .001	0.774
n2_calib_std	-	n3_calib_std	0.766	23	0.451	0.156

			Wilcoxon		
			W	p	Rank-Biserial Correlation
n1_calib_min	-	n2_calib_min	1.000	< .001	-0.993
	-	n3_calib_min	35.000	< .001	-0.767
n2_calib_min	-	n3_calib_min	176.000	0.473	0.173
n1_calib_max	-	n2_calib_max	54.000	0.005	-0.640
	-	n3_calib_max	42.000	0.001	-0.720
n2_calib_max	-	n3_calib_max	161.000	0.768	0.073
n1_calib_quantile75	-	n2_calib_quantile75	4.000	< .001	-0.973
	-	n3_calib_quantile75	0.000	< .001	-1.000
n2_calib_quantile75	-	n3_calib_quantile75	144.000	0.877	-0.040

			T-test			
			t	df	p	Cohen's d
n1_mean	-	n2_mean	-11.901	23	< .001	-2.429
	-	n3_mean	-10.107	23	< .001	-2.063
n2_mean	-	n3_mean	-1.717	23	0.099	-0.350
n1_median	-	n2_median	-9.339	23	< .001	-1.906
	-	n3_median	-9.173	23	< .001	-1.872
n2_median	-	n3_median	-1.341	23	0.193	-0.274
n1_max	-	n2_max	-6.011	23	< .001	-1.227
	-	n3_max	-6.416	23	< .001	-1.310
n2_max	-	n3_max	-2.717	23	0.012	-0.555
n1_quantile25	-	n2_quantile25	-12.125	23	< .001	-2.475
	-	n3_quantile25	-9.713	23	< .001	-1.983
n2_quantile25	-	n3_quantile25	-1.873	23	0.074	-0.382

			Wilcoxon		
			W	p	Rank-Biserial Correlation
n1_min	-	n2_min	0.000	< .001	-1.000
	-	n3_min	30.000	< .001	-0.800
n2_min	-	n3_min	166.000	0.663	0.107
n1_P2P	-	n2_P2P	297.000	< .001	0.980
	-	n3_P2P	243.000	0.007	0.620
n2_P2P	-	n3_P2P	101.000	0.169	-0.327
n1_quantile75	-	n2_quantile75	0.000	< .001	-1.000
	-	n3_quantile75	0.000	< .001	-1.000
n2_quantile75	-	n3_quantile75	101.000	0.169	-0.327
n1_IQR	-	n2_IQR	295.000	< .001	0.967
	-	n3_IQR	290.000	< .001	0.933
n2_IQR	-	n3_IQR	174.000	0.509	0.160
n1_std	-	n2_std	296.000	< .001	0.973
	-	n3_std	288.000	< .001	0.920
n2_std	-	n3_std	140.000	0.790	-0.067

Table A14. Post hoc tests: on the left using 0-back calibration, on the right without using 0-back calibration

C EDA ANALYSIS STUDY 1

C.1 Descriptive Statistics

	N	Mean	SD	SE
n1_gsr_mean	24	-0.331	1.071	0.219
n2_gsr_mean	24	0.240	0.947	0.193
n3_gsr_mean	24	-0.004	0.765	0.156
n1_gsr_var	24	1.096	1.203	0.246
n2_gsr_var	24	0.895	0.808	0.165
n3_gsr_var	24	1.060	1.229	0.251
n1_gsr_peak_mean	24	0.017	1.144	0.234
n2_gsr_peak_mean	24	0.528	0.959	0.196
n3_gsr_peak_mean	24	0.269	0.885	0.181
n1_gsr_peak_var	24	1.566	1.935	0.395
n2_gsr_peak_var	24	1.120	1.015	0.207
n3_gsr_peak_var	24	1.356	1.667	0.340
n1_gsr_max_peak	24	2.176	2.054	0.419
n2_gsr_max_peak	24	2.607	1.810	0.369
n3_gsr_max_peak	24	2.229	1.698	0.347
n1_peak_count	24	23.917	7.027	1.434
n2_peak_count	24	27.333	5.983	1.221
n3_peak_count	24	24.250	7.958	1.624

Table A15. Descriptives EDA measures (without 0-back calibration)

	N	Mean	SD	SE
n1_calib_gsr_mean	24	-0.017	0.034	0.007
n2_calib_gsr_mean	24	-0.005	0.042	0.009
n3_calib_gsr_mean	24	-0.015	0.045	0.009
n1_calib_gsr_var	24	0.005	0.008	0.002
n2_calib_gsr_var	24	0.005	0.006	0.001
n3_calib_gsr_var	24	0.010	0.020	0.004
n1_calib_gsr_peak_mean	24	0.008	0.042	0.009
n2_calib_gsr_peak_mean	24	0.017	0.041	0.008
n3_calib_gsr_peak_mean	24	0.012	0.050	0.010
n1_calib_gsr_peak_var	24	0.006	0.009	0.002
n2_calib_gsr_peak_var	24	0.006	0.007	0.001
n3_calib_gsr_peak_var	24	0.011	0.021	0.004
n1_calib_gsr_max_peak	24	0.144	0.118	0.024
n2_calib_gsr_max_peak	24	0.156	0.123	0.025
n3_calib_gsr_max_peak	24	0.171	0.154	0.031
n1_calib_peak_count	24	24.417	6.613	1.350
n2_calib_peak_count	24	27.542	5.823	1.189
n3_calib_peak_count	24	24.625	7.751	1.582

Table A16. Descriptives EDA measures (using 0-back calibration)

C.2 ANOVA and Friedman Test

Repeated Measures ANOVA

	Sum of Squares	df	Mean Square	F	p	η^2	${\eta_p}^2$
VAR	0.555	2	0.278	0.468	0.629	0.020	0.020
PEAK MEAN	3.138	2	1.569	1.118	0.336	0.046	0.046
PEAK COUNT	170.333	2	85.167	3.805	0.030	0.142	0.142

Friedman Test

	df	Chi-Squared	p	Kendall's W
MEAN	2	1.083	0.582	0.031
VAR	2	0.583	0.747	0.760
PEAK VAR	2	1.083	0.582	0.780
PEAK MAX	2	0.083	0.959	0.479
PEAK COUNT	2	2.957	0.228	0.693

Table A17. without n0 calibration

Repeated Measures ANOVA

	Sum of Squares	df	Mean Square	F	p	η^2	${\eta_p}^2$
PEAK MEAN	8.617e-4	2	4.309e-4	0.361	0.699	0.015	0.015
PEAK COUNT	146.528	2	73.264	3.448	0.040	0.130	0.130

Friedman Test

	df	Chi-Squared	p	Kendall's W
MEAN	2	0.333	0.846	0.585
VAR	2	1.083	0.582	0.633
PEAK VAR	2	1.083	0.582	0.780
PEAK COUNT	2	2.716	0.257	0.680
PEAK MAX	2	0.250	0.882	0.585

Table A18. using 0-back calibration

C.3 Assumption Checks

Without 0-back calibration				
	Mauchly's W	p	Greenhouse-Geisser ϵ	Huynh-Feldt ϵ
MEAN	0.839	0.146	0.862	0.925
VAR	0.910	0.354	0.917	0.993
PEAK MEAN	0.831	0.131	0.856	0.918
PEAK VAR	0.968	0.700	0.969	1.000
PEAK COUNT	0.899	0.310	0.908	0.982
PEAK MAX	0.911	0.361	0.919	0.995

Using 0-back calibration				
	Mauchly's W	p	Greenhouse-Geisser ϵ	Huynh-Feldt ϵ
MEAN	0.933	0.465	0.937	1.000
VAR	0.263	$< .001$	0.576	0.587
PEAK MEAN	0.926	0.430	0.931	1.000
PEAK VAR	0.968	0.700	0.969	1.000
PEAK COUNT	0.921	0.406	0.927	1.000
PEAK MAX	0.819	0.111	0.847	0.907

Table A19. Mauchly's Test of Spericity

			W	p
n1_gsr_mean	-	n2_gsr_mean	0.914	0.042
	-	n3_gsr_mean	0.984	0.958
n2_gsr_mean	-	n3_gsr_mean	0.963	0.512
n1_gsr_var	-	n2_gsr_var	0.859	0.003
	-	n3_gsr_var	0.929	0.095
n2_gsr_var	-	n3_gsr_var	0.857	0.003
n1_gsr_peak_mean	-	n2_gsr_peak_mean	0.929	0.093
	-	n3_gsr_peak_mean	0.979	0.873
n2_gsr_peak_mean	-	n3_gsr_peak_mean	0.964	0.535
n1_gsr_peak_var	-	n2_gsr_peak_var	0.778	< .001
	-	n3_gsr_peak_var	0.917	0.051
n2_gsr_peak_var	-	n3_gsr_peak_var	0.805	< .001
n1_gsr_max_peak	-	n2_gsr_max_peak	0.911	0.038
	-	n3_gsr_max_peak	0.972	0.719
n2_gsr_max_peak	-	n3_gsr_max_peak	0.982	0.927
n1_peak_count	-	n2_peak_count	0.972	0.716
	-	n3_peak_count	0.976	0.802
n2_peak_count	-	n3_peak_count	0.968	0.609

Table A20. Test of Normality (Shapiro-Wilk): without using 0-back calibration

			W	p
n1_calib_gsr_mean	-	n2_calib_gsr_mean	0.948	0.251
	-	n3_calib_gsr_mean	0.981	0.920
n2_calib_gsr_mean	-	n3_calib_gsr_mean	0.973	0.741
n1_calib_gsr_var	-	n2_calib_gsr_var	0.724	< .001
	-	n3_calib_gsr_var	0.611	< .001
n2_calib_gsr_var	-	n3_calib_gsr_var	0.592	< .001
n1_calib_gsr_peak_mean	-	n2_calib_gsr_peak_mean	0.974	0.775
	-	n3_calib_gsr_peak_mean	0.913	0.041
n2_calib_gsr_peak_mean	-	n3_calib_gsr_peak_mean	0.901	0.022
n1_calib_gsr_peak_var	-	n2_calib_gsr_peak_var	0.726	< .001
	-	n3_calib_gsr_peak_var	0.659	< .001
n2_calib_gsr_peak_var	-	n3_calib_gsr_peak_var	0.662	< .001
n1_calib_gsr_max_peak	-	n2_calib_gsr_max_peak	0.985	0.965
	-	n3_calib_gsr_max_peak	0.913	0.042
n2_calib_gsr_max_peak	-	n3_calib_gsr_max_peak	0.888	0.012
n1_calib_peak_count	-	n2_calib_peak_count	0.970	0.668
	-	n3_calib_peak_count	0.977	0.838
n2_calib_peak_count	-	n3_calib_peak_count	0.959	0.415

Table A21. Test of Normality (Shapiro-Wilk): using 0-back calibration

C.4 Post Hoc Tests

Table A22. Paired Samples T-Test

			Test	Statistic	df	p	Effect Size
n1_peak_count	-	n2_peak_count	T-test	-2.228	23	0.036	-0.455
			Wilcoxon	70.000		0.040	-0.533
n1_peak_count	-	n3_peak_count	T-test	-0.289	23	0.775	-0.059
			Wilcoxon	125.000		0.703	-0.167
n2_peak_count	-	n3_peak_count	T-test	2.227	23	0.036	0.455
			Wilcoxon	219.500		0.048	0.463

D BVP ANALYSIS STUDY 1

D.1 Descriptives

	N	Mean	SD	SE
n1_hr_bpm	24	74.737	9.721	1.984
n2_hr_bpm	24	76.431	10.997	2.245
n3_hr_bpm	24	76.166	10.090	2.060
n1_ibi_s	24	0.820	0.110	0.022
n2_ibi_s	24	0.804	0.113	0.023
n3_ibi_s	24	0.806	0.108	0.022
n1_ibi_variability	24	-3.953e-5	7.448e-4	1.520e-4
n2_ibi_variability	24	-2.746e-4	7.899e-4	1.612e-4
n3_ibi_variability	24	-1.086e-4	0.001	2.458e-4
n1_ibi_variability_std	24	0.058	0.022	0.005
n2_ibi_variability_std	24	0.056	0.022	0.005
n3_ibi_variability_std	24	0.058	0.022	0.004
n1_hr_variability	24	-0.003	0.062	0.013
n2_hr_variability	24	0.027	0.067	0.014
n3_hr_variability	24	0.013	0.099	0.020
n1_hr_variability_std	24	5.164	1.743	0.356
n2_hr_variability_std	24	5.198	1.793	0.366
n3_hr_variability_std	24	5.287	1.682	0.343
n1_ibiv_squared	24	0.058	0.022	0.004
n2_ibiv_squared	24	0.056	0.022	0.004
n3_ibiv_squared	24	0.057	0.022	0.004
n1_hrv_squared	24	5.154	1.736	0.354
n2_hrv_squared	24	5.187	1.785	0.364
n3_hrv_squared	24	5.277	1.677	0.342

Table A23. Descriptives traditional HR measures

D.2 ANOVA and Friedman Test

Repeated Measures ANOVA							
	Sum of Squares	df	Mean Square	F	p	η^2	${\eta_p}^2$
IBI	0.004	2	0.002	2.744	0.075	0.107	0.107
HR	39.841	2	19.920	2.493	0.094	0.098	0.098
HRVSD	0.191	2	0.096	0.111	0.895	0.005	0.005

Friedman Test				
	df	Chi-Squared	p	Kendall's W
HRV	2	1.333	0.513	0.270
IBIV	2	2.333	0.311	0.264
IBIVSD	2	0.000	1.000	0.847
RMSIBI	2	0.000	1.000	0.847

Table A25. Traditional Measures

Friedman Test				
	df	Chi-Squared	p	Kendall's W
MAX PEAK	2	0.083	0.959	0.194
PEAK COUNT	2	0.189	0.910	0.847
BVP MEAN	2	0.083	0.959	0.030
BVP VAR	2	2.083	0.353	0.065
BVP STD	2	2.083	0.353	0.065
PEAK VAR	2	1.583	0.453	0.140
PEAK MEAN	2	1.750	0.417	0.496
PEAK STD	2	1.583	0.453	0.140

Table A26. BVP Measures

D.3 Assumption Checks

D.3.1 Sphericity

Traditional Measures

	Mauchly's W	p	Greenhouse-Geisser ϵ	Huynh-Feldt ϵ
IBI	0.898	0.307	0.908	0.981
HR	0.953	0.592	0.956	1.000
HRV	0.673	0.013	0.754	0.795
IBIV	0.598	0.004	0.713	0.747
HRVSD	0.888	0.272	0.900	0.971
IBIVSD	0.977	0.774	0.977	1.000
RMSIBI	0.977	0.773	0.977	1.000

BVP Measures

	Mauchly's W	p	Greenhouse-Geisser ϵ	Huynh-Feldt ϵ
MAX PEAK	0.768	0.055	0.812	0.864
PEAK COUNT	0.843	0.153	0.865	0.928
BVP MEAN	0.735	0.034	0.790	0.839
BVP VAR	0.920	0.398	0.926	1.000
BVP STD	0.915	0.376	0.922	0.998
PEAK VAR	0.833	0.134	0.857	0.919
PEAK MEAN	0.802	0.088	0.835	0.892
PEAK STD	0.835	0.137	0.858	0.921

Table A27. Mauchly's Test of Spericity

D.3.2 Normality

			W	p
n1_hr_bpm	-	n2_hr_bpm	0.949	0.256
	-	n3_hr_bpm	0.941	0.175
n2_hr_bpm	-	n3_hr_bpm	0.978	0.856
n1_ibi_s	-	n2_ibi_s	0.974	0.773
	-	n3_ibi_s	0.950	0.266
n2_ibi_s	-	n3_ibi_s	0.978	0.854
n1_ibi_variability	-	n2_ibi_variability	0.967	0.594
	-	n3_ibi_variability	0.897	0.019
n2_ibi_variability	-	n3_ibi_variability	0.918	0.053
n1_ibi_variability_std	-	n2_ibi_variability_std	0.979	0.877
	-	n3_ibi_variability_std	0.891	0.014
n2_ibi_variability_std	-	n3_ibi_variability_std	0.917	0.050
n1_hr_variability	-	n2_hr_variability	0.969	0.654
	-	n3_hr_variability	0.975	0.783
n2_hr_variability	-	n3_hr_variability	0.951	0.285
n1_hr_variability_std	-	n2_hr_variability_std	0.978	0.863
	-	n3_hr_variability_std	0.969	0.634
n2_hr_variability_std	-	n3_hr_variability_std	0.962	0.478
n1_ibiv_squared	-	n2_ibiv_squared	0.979	0.878
	-	n3_ibiv_squared	0.892	0.015
n2_ibiv_squared	-	n3_ibiv_squared	0.917	0.051
n1_hrv_squared	-	n2_hrv_squared	0.978	0.864
	-	n3_hrv_squared	0.969	0.639
n2_hrv_squared	-	n3_hrv_squared	0.962	0.472

Table A28. Test of Normality (Shapiro-Wilk): Traditional Measures

			W	p
bvp_mean_n1	-	bvp_mean_n2	0.953	0.307
	-	bvp_mean_n3	0.971	0.701
bvp_mean_n2	-	bvp_mean_n3	0.929	0.094
bvp_std_n1	-	bvp_std_n2	0.971	0.692
	-	bvp_std_n3	0.926	0.078
bvp_std_n2	-	bvp_std_n3	0.854	0.003
bvp_var_n1	-	bvp_var_n2	0.963	0.510
	-	bvp_var_n3	0.919	0.056
bvp_var_n2	-	bvp_var_n3	0.877	0.007
bvp_peak_mean_n1	-	bvp_peak_mean_n2	0.939	0.151
	-	bvp_peak_mean_n3	0.871	0.006
bvp_peak_mean_n2	-	bvp_peak_mean_n3	0.849	0.002
bvp_peak_var_n1	-	bvp_peak_var_n2	0.936	0.130
	-	bvp_peak_var_n3	0.895	0.017
bvp_peak_var_n2	-	bvp_peak_var_n3	0.729	< .001
bvp_max_peak_n1	-	bvp_max_peak_n2	0.881	0.009
	-	bvp_max_peak_n3	0.864	0.004
bvp_max_peak_n2	-	bvp_max_peak_n3	0.935	0.128
bvp_peak_std_n1	-	bvp_peak_std_n2	0.967	0.604
	-	bvp_peak_std_n3	0.917	0.050
bvp_peak_std_n2	-	bvp_peak_std_n3	0.749	< .001
bvp_peak_count_n1	-	bvp_peak_count_n2	0.973	0.732
	-	bvp_peak_count_n3	0.899	0.021
bvp_peak_count_n2	-	bvp_peak_count_n3	0.956	0.356

Table A29. Test of Normality (Shapiro-Wilk): BVP Measures

	N	Mean	SD	SE
bvp_mean_n1	24	0.004	0.026	0.005
bvp_mean_n2	24	-0.002	0.027	0.006
bvp_mean_n3	24	-0.002	0.017	0.003
bvp_std_n1	24	0.982	0.328	0.067
bvp_std_n2	24	0.933	0.281	0.057
bvp_std_n3	24	0.956	0.267	0.055
bvp_var_n1	24	1.069	0.641	0.131
bvp_var_n2	24	0.947	0.500	0.102
bvp_var_n3	24	0.983	0.570	0.116
bvp_peak_mean_n1	24	0.408	0.162	0.033
bvp_peak_mean_n2	24	0.468	0.276	0.056
bvp_peak_mean_n3	24	0.419	0.170	0.035
bvp_peak_var_n1	24	0.943	0.620	0.126
bvp_peak_var_n2	24	0.829	0.420	0.086
bvp_peak_var_n3	24	0.960	0.672	0.137
bvp_max_peak_n1	24	5.022	3.547	0.724
bvp_max_peak_n2	24	4.164	1.974	0.403
bvp_max_peak_n3	24	4.380	1.898	0.387
bvp_peak_std_n1	24	0.915	0.332	0.068
bvp_peak_std_n2	24	0.874	0.261	0.053
bvp_peak_std_n3	24	0.933	0.305	0.062
bvp_peak_count_n1	24	143.875	43.249	8.828
bvp_peak_count_n2	24	136.458	35.743	7.296
bvp_peak_count_n3	24	139.375	35.038	7.152

Table A24. Descriptives BVP measures

E COGNITIVE STATE CLASSIFICATION STUDY 1

E.1 Scores of all pipelines

win size	classes	model	scaling approach	feature approach	acc mean	acc sd	class sizes	confusion matrix
WIN60	1v3	RFC50	SCALED	F27	0.901	0.11	[95, 91]	[84 11] [7 84]
WIN60	1v2	RFC50	SCALED	F27	0.891	0.085	[95, 96]	[83 12] [9 87]
WIN60	1v2	RFC10	SCALED	F27	0.885	0.09	[95, 96]	[84 11] [11 85]
WIN60	1v3	RFC50	SCALED	F9	0.885	0.11	[95, 91]	[85 10] [11 80]
WIN60	1v2	RFC50	SCALED	F9	0.879	0.107	[95, 96]	[81 14] [9 87]
WIN60	1v3	RFC10	SCALED	F27	0.875	0.133	[95, 91]	[83 12] [11 80]
WIN60	1v3	RFC10	SCALED	F9	0.869	0.107	[95, 91]	[84 11] [13 78]
WIN30	1v3	RFC50	SCALED	F27	0.867	0.121	[3319, 3209]	[2875 444] [402 2807]
WIN30	1v2	RFC50	SCALED	F27	0.86	0.094	[3319, 3347]	[2824 495] [445 2902]
WIN30	1v3	RFC10	SCALED	F27	0.859	0.12	[3319, 3209]	[2901 418] [486 2723]
WIN10	1v3	RFC50	SCALED	F27	0.857	0.119	[5219, 5049]	[4422 797] [649 4400]
WIN60	1v2	RFC10	SCALED	F9	0.854	0.131	[95, 96]	[80 15] [13 83]
WIN30	1v2	RFC10	SCALED	F27	0.854	0.089	[3319, 3347]	[2865 454] [521 2826]
WIN60	1v3	RFC50	CALIB	F9	0.847	0.157	[95, 91]	[84 11] [18 73]
WIN30	1v3	RFC50	CALIB	F27	0.847	0.109	[3319, 3209]	[2920 399] [599 2610]
WIN10	1v3	RFC10	SCALED	F27	0.846	0.12	[5219, 5049]	[4436 783] [773 4276]
WIN5	1v3	RFC50	SCALED	F27	0.844	0.122	[5694, 5509]	[4729 965] [753 4756]
WIN60	1v2	RFC10	CALIB	F27	0.843	0.148	[95, 96]	[85 10] [20 76]
WIN60	1v3	RFC50	CALIB	F27	0.842	0.167	[95, 91]	[81 14] [16 75]
WIN60	1v3	RFC10	CALIB	F9	0.842	0.163	[95, 91]	[85 10] [20 71]
WIN10	1v3	RFC50	SCALED	F9	0.841	0.119	[5219, 5049]	[4354 865] [740 4309]
WIN30	1v3	RFC50	SCALED	F9	0.84	0.112	[3319, 3209]	[2857 462] [573 2636]
WIN30	1v3	RFC10	CALIB	F27	0.839	0.117	[3319, 3209]	[2951 368] [678 2531]
WIN10	1v2	RFC50	SCALED	F27	0.838	0.098	[5219, 5267]	[4217 1002] [705 4562]
WIN3	1v3	RFC50	SCALED	F27	0.836	0.12	[5884, 5693]	[4829 1055] [805 4888]
WIN10	1v2	RFC10	SCALED	F27	0.835	0.093	[5219, 5267]	[4301 918] [821 4446]
WIN30	1v3	RFC50	CALIB	F9	0.834	0.112	[3319, 3209]	[2902 417] [669 2540]
WIN5	1v3	RFC10	SCALED	F27	0.832	0.122	[5694, 5509]	[4760 934] [924 4585]
WIN5	1v3	RFC50	SCALED	F9	0.831	0.115	[5694, 5509]	[4661 1033] [830 4679]
WIN10	1v3	RFC10	SCALED	F9	0.831	0.111	[5219, 5049]	[4392 827] [893 4156]
WIN30	1v3	RFC10	SCALED	F9	0.829	0.12	[3319, 3209]	[2860 459] [644 2565]
WIN3	1v3	RFC50	SCALED	F9	0.828	0.116	[5884, 5693]	[4826 1058] [908 4785]
WIN5	1v2	RFC50	SCALED	F27	0.826	0.094	[5694, 5747]	[4528 1166] [834 4913]
WIN30	1v3	RFC10	CALIB	F9	0.825	0.109	[3319, 3209]	[2891 428] [708 2501]
WIN3	1v3	RFC10	SCALED	F27	0.824	0.118	[5884, 5693]	[4900 984] [1024 4669]
WIN30	1v2	RFC50	SCALED	F9	0.822	0.095	[3319, 3347]	[2709 610] [581 2766]
WIN10	1v2	RFC50	SCALED	F9	0.821	0.092	[5219, 5267]	[4199 1020] [866 4401]
WIN60	1v3	RFC10	CALIB	F27	0.821	0.157	[95, 91]	[82 13] [20 71]
WIN5	1v3	RFC10	SCALED	F9	0.82	0.111	[5694, 5509]	[4712 982] [1007 4502]
WIN5	1v2	RFC10	SCALED	F27	0.819	0.089	[5694, 5747]	[4644 1050] [1033 4714]
WIN3	1v3	RFC10	SCALED	F9	0.818	0.11	[5884, 5693]	[4867 1017] [1062 4631]
WIN5	1v2	RFC50	SCALED	F9	0.815	0.093	[5694, 5747]	[4498 1196] [929 4818]
WIN3	1v2	RFC50	SCALED	F27	0.815	0.094	[5884, 5939]	[4630 1254] [939 5000]
WIN30	1v2	RFC10	SCALED	F9	0.813	0.095	[3319, 3347]	[2750 569] [681 2666]
WIN60	1v2	RFC50	CALIB	F27	0.812	0.173	[95, 96]	[80 15] [21 75]
WIN10	1v2	RFC10	SCALED	F9	0.809	0.086	[5219, 5267]	[4240 979] [1031 4236]
WIN10	1v3	RFC50	CALIB	F27	0.806	0.096	[5219, 5049]	[4321 898] [1070 3979]
WIN5	1v2	RFC10	SCALED	F9	0.803	0.087	[5694, 5747]	[4568 1126] [1132 4615]
WIN3	1v2	RFC50	SCALED	F9	0.803	0.09	[5884, 5939]	[4598 1286] [1053 4886]
WIN3	1v2	RFC10	SCALED	F27	0.801	0.093	[5884, 5939]	[4707 1177] [1182 4757]
WIN10	1v3	RFC10	CALIB	F27	0.796	0.098	[5219, 5049]	[4403 816] [1266 3783]
WIN60	1v3	RFC50	RAW	F27	0.795	0.15	[95, 91]	[77 18] [20 71]
WIN5	1v3	RFC50	CALIB	F27	0.791	0.098	[5694, 5509]	[4661 1033] [1292 4217]
WIN60	1v2	RFC50	RAW	F27	0.791	0.131	[95, 96]	[78 17] [23 73]
WIN3	1v2	RFC10	SCALED	F9	0.791	0.086	[5884, 5939]	[4673 1211] [1271 4668]
WIN60	1v2	RFC10	RAW	F27	0.786	0.158	[95, 96]	[79 16] [25 71]
WIN60	1v2	RFC50	CALIB	F9	0.786	0.175	[95, 96]	[76 19] [22 74]
WIN10	1v3	RFC50	CALIB	F9	0.784	0.09	[5219, 5049]	[4232 987] [1220 3829]
WIN60	1v2	RFC10	CALIB	F9	0.781	0.157	[95, 96]	[77 18] [24 72]
WIN3	1v3	RFC50	CALIB	F27	0.781	0.097	[5884, 5693]	[4762 1122] [1387 4306]

Table A30. Cross validation results: all aproaches 0

win size	classes	model	scaling approach	feature approach	acc mean	acc sd	class sizes	confusion matrix
WIN5	1v3	RFC10	CALIB	F27	0.778	0.094	[5694, 5509]	[4707 987] [1490 4019]
WIN10	1v3	RFC10	CALIB	F9	0.777	0.092	[5219, 5049]	[4288 931] [1353 3696]
WIN30	1v2	RFC10	CALIB	F27	0.774	0.144	[3319, 3347]	[2854 465] [1043 2304]
WIN5	1v3	RFC50	CALIB	F9	0.771	0.092	[5694, 5509]	[4560 1134] [1405 4104]
WIN30	1v2	RFC50	CALIB	F27	0.771	0.143	[3319, 3347]	[2810 509] [1018 2329]
WIN3	1v3	RFC10	CALIB	F27	0.769	0.097	[5884, 5693]	[4815 1069] [1584 4109]
WIN3	1v3	RFC50	CALIB	F9	0.764	0.095	[5884, 5693]	[4625 1259] [1442 4251]
WIN5	1v3	RFC10	CALIB	F9	0.761	0.095	[5694, 5509]	[4623 1071] [1580 3929]
WIN10	1v2	RFC50	CALIB	F27	0.759	0.113	[5219, 5267]	[4274 945] [1578 3689]
WIN3	1v3	RFC10	CALIB	F9	0.755	0.096	[5884, 5693]	[4704 1180] [1627 4066]
WIN30	1v2	RFC50	CALIB	F9	0.752	0.135	[3319, 3347]	[2756 563] [1090 2257]
WIN10	1v2	RFC10	CALIB	F27	0.748	0.114	[5219, 5267]	[4304 915] [1718 3549]
WIN30	1v2	RFC10	CALIB	F9	0.744	0.138	[3319, 3347]	[2795 524] [1183 2164]
WIN60	1v3	RFC10	RAW	F27	0.743	0.169	[95, 91]	[73 22] [26 65]
WIN5	1v2	RFC50	CALIB	F27	0.739	0.112	[5694, 5747]	[4446 1248] [1730 4017]
WIN3	1v2	RFC50	CALIB	F27	0.73	0.112	[5884, 5939]	[4512 1372] [1820 4119]
WIN5	1v2	RFC10	CALIB	F27	0.73	0.119	[5694, 5747]	[4504 1190] [1895 3852]
WIN10	1v2	RFC50	CALIB	F9	0.723	0.115	[5219, 5267]	[4070 1149] [1750 3517]
WIN10	1v2	RFC10	CALIB	F9	0.722	0.105	[5219, 5267]	[4133 1086] [1820 3447]
WIN5	1v2	RFC50	CALIB	F9	0.717	0.104	[5694, 5747]	[4331 1363] [1866 3881]
WIN60	1v2	RFC50	RAW	F9	0.714	0.167	[95, 96]	[71 24] [31 65]
WIN3	1v2	RFC10	CALIB	F27	0.712	0.11	[5884, 5939]	[4560 1324] [2072 3867]
WIN3	1v2	RFC50	CALIB	F9	0.71	0.113	[5884, 5939]	[4418 1466] [1958 3981]
WIN5	1v2	RFC10	CALIB	F9	0.71	0.109	[5694, 5747]	[4403 1291] [2017 3730]
WIN3	1v2	RFC10	CALIB	F9	0.707	0.11	[5884, 5939]	[4532 1352] [2104 3835]
WIN30	1v3	RFC10	RAW	F27	0.707	0.148	[3319, 3209]	[2663 656] [1268 1941]
WIN60	1v2	RFC10	RAW	F9	0.693	0.184	[95, 96]	[70 25] [34 62]
WIN30	1v2	RFC50	RAW	F9	0.686	0.134	[3319, 3347]	[2359 960] [1141 2206]
WIN30	1v3	RFC50	RAW	F27	0.683	0.162	[3319, 3209]	[2502 817] [1257 1952]
WIN10	1v2	RFC50	RAW	F27	0.682	0.143	[5219, 5267]	[3706 1513] [1843 3424]
WIN60	1v3	RFC10	RAW	F9	0.681	0.156	[95, 91]	[75 20] [40 51]
WIN30	1v2	RFC50	RAW	F27	0.68	0.151	[3319, 3347]	[2479 840] [1301 2046]
WIN30	1v2	RFC10	RAW	F27	0.676	0.117	[3319, 3347]	[2522 797] [1366 1981]
WIN5	1v2	RFC50	RAW	F27	0.672	0.137	[5694, 5747]	[3986 1708] [2058 3689]
WIN10	1v3	RFC50	RAW	F27	0.67	0.127	[5219, 5049]	[3755 1464] [1944 3105]
WIN60	1v3	RFC50	RAW	F9	0.67	0.158	[95, 91]	[71 24] [37 54]
WIN5	1v3	RFC50	RAW	F27	0.67	0.114	[5694, 5509]	[4024 1670] [2040 3469]
WIN30	1v2	RFC10	RAW	F9	0.667	0.125	[3319, 3347]	[2342 977] [1255 2092]
WIN10	1v2	RFC10	RAW	F27	0.664	0.142	[5219, 5267]	[3783 1436] [2102 3165]
WIN5	1v2	RFC10	RAW	F27	0.662	0.131	[5694, 5747]	[4139 1555] [2333 3414]
WIN3	1v2	RFC50	RAW	F27	0.661	0.131	[5884, 5939]	[4018 1866] [2162 3777]
WIN5	1v3	RFC10	RAW	F27	0.66	0.112	[5694, 5509]	[4143 1551] [2274 3235]
WIN3	1v3	RFC50	RAW	F27	0.656	0.115	[5884, 5693]	[4090 1794] [2204 3489]
WIN30	1v3	RFC10	RAW	F9	0.656	0.137	[3319, 3209]	[2440 879] [1379 1830]
WIN10	1v2	RFC50	RAW	F9	0.655	0.13	[5219, 5267]	[3529 1690] [1938 3329]
WIN3	1v2	RFC10	RAW	F27	0.652	0.117	[5884, 5939]	[4201 1683] [2447 3492]
WIN30	1v3	RFC50	RAW	F9	0.649	0.138	[3319, 3209]	[2329 990] [1302 1907]
WIN10	1v3	RFC10	RAW	F27	0.648	0.121	[5219, 5049]	[3794 1425] [2201 2848]
WIN10	1v2	RFC10	RAW	F9	0.646	0.119	[5219, 5267]	[3624 1595] [2128 3139]
WIN3	1v3	RFC10	RAW	F27	0.644	0.106	[5884, 5693]	[4212 1672] [2468 3225]
WIN5	1v2	RFC50	RAW	F9	0.641	0.114	[5694, 5747]	[3730 1964] [2153 3594]
WIN5	1v3	RFC50	RAW	F9	0.639	0.094	[5694, 5509]	[3839 1855] [2176 3333]
WIN5	1v3	RFC10	RAW	F9	0.635	0.092	[5694, 5509]	[3976 1718] [2361 3148]
WIN10	1v3	RFC50	RAW	F9	0.635	0.105	[5219, 5049]	[3566 1653] [2093 2956]
WIN3	1v2	RFC50	RAW	F9	0.632	0.108	[5884, 5939]	[3777 2107] [2256 3683]
WIN5	1v2	RFC10	RAW	F9	0.632	0.099	[5694, 5747]	[3877 1817] [2401 3346]
WIN10	1v3	RFC10	RAW	F9	0.623	0.095	[5219, 5049]	[3607 1612] [2250 2799]
WIN30	1v2v3	RFC50	SCALED	F27	0.62	0.1	[3319, 3347, 3209]	[2745 340 234] [333 1755 1259] [265 1327 1617]
WIN3	1v2	RFC10	RAW	F9	0.618	0.099	[5884, 5939]	[3888 1996] [2529 3410]
WIN3	1v3	RFC50	RAW	F9	0.617	0.097	[5884, 5693]	[3841 2043] [2383 3310]

Table A31. Cross validation results: all aproaches 1

win size	classes	model	scaling approach	feature approach	acc mean	acc sd	class sizes	confusion matrix
WIN60	1v2v3	RFC50	SCALED	F27	0.613	0.138	[95, 96, 91]	[80 8 7] [8 49 39] [5 43 43]
WIN3	1v3	RFC10	RAW	F9	0.613	0.094	[5884, 5693]	[3992 1892] [2575 3118]
WIN30	1v2v3	RFC10	SCALED	F27	0.606	0.096	[3319, 3347, 3209]	[2696 360 263] [394 1820 1133] [350 1396 1463]
WIN60	1v2v3	RFC10	SCALED	F27	0.593	0.142	[95, 96, 91]	[79 10 6] [8 46 42] [6 43 42]
WIN60	1v2v3	RFC50	SCALED	F9	0.589	0.148	[95, 96, 91]	[77 11 7] [7 43 46] [4 42 45]
WIN60	1v2v3	RFC10	SCALED	F9	0.589	0.147	[95, 96, 91]	[77 10 8] [11 45 40] [3 45 43]
WIN60	2v3	RFC10	RAW	F27	0.583	0.2	[96, 91]	[58 38] [39 52]
WIN30	1v2v3	RFC50	SCALED	F9	0.575	0.093	[3319, 3347, 3209]	[2650 428 241] [451 1570 1326] [353 1411 1445]
WIN30	1v2v3	RFC10	SCALED	F9	0.574	0.091	[3319, 3347, 3209]	[2674 405 240] [503 1640 1204] [415 1458 1336]
WIN10	1v2v3	RFC50	SCALED	F27	0.571	0.082	[5219, 5267, 5049]	[4129 715 375] [574 2415 2278] [433 2292 2324]
WIN5	1v2v3	RFC50	SCALED	F27	0.562	0.081	[5694, 5747, 5509]	[4415 779 500] [707 2588 2452] [522 2455 2532]
WIN60	2v3	RFC50	RAW	F27	0.561	0.19	[96, 91]	[58 38] [43 48]
WIN10	1v2v3	RFC10	SCALED	F27	0.56	0.073	[5219, 5267, 5049]	[4109 752 358] [684 2572 2011] [491 2541 2017]
WIN30	2v3	RFC50	SCALED	F27	0.556	0.128	[3347, 3209]	[1909 1438] [1503 1706]
WIN5	1v2v3	RFC10	SCALED	F27	0.555	0.071	[5694, 5747, 5509]	[4402 825 467] [765 2729 2253] [583 2665 2261]
WIN3	1v2v3	RFC50	SCALED	F27	0.555	0.072	[5884, 5939, 5693]	[4533 856 495] [778 2634 2527] [546 2595 2552]
WIN10	1v2v3	RFC10	SCALED	F9	0.553	0.067	[5219, 5267, 5049]	[4100 704 415] [748 2423 2096] [558 2435 2056]
WIN5	1v2v3	RFC50	SCALED	F9	0.551	0.069	[5694, 5747, 5509]	[4367 754 573] [763 2518 2466] [535 2528 2446]
WIN10	1v2v3	RFC50	SCALED	F9	0.549	0.076	[5219, 5267, 5049]	[4078 693 448] [679 2222 2366] [532 2274 2243]
WIN3	1v2v3	RFC50	SCALED	F9	0.548	0.066	[5884, 5939, 5693]	[4436 886 562] [863 2590 2486] [586 2521 2586]
WIN30	1v2v3	RFC50	CALIB	F27	0.548	0.084	[3319, 3347, 3209]	[2739 316 264] [794 1297 1256] [455 1365 1389]
WIN5	1v2v3	RFC10	SCALED	F9	0.547	0.062	[5694, 5747, 5509]	[4357 808 529] [849 2661 2237] [608 2651 2250]
WIN60	1v2v3	RFC50	RAW	F27	0.542	0.155	[95, 96, 91]	[69 14 12] [21 44 31] [22 29 40]
WIN3	1v2v3	RFC10	SCALED	F27	0.542	0.06	[5884, 5939, 5693]	[4523 874 487] [868 2765 2306] [634 2866 2193]
WIN30	2v3	RFC50	SCALED	F9	0.54	0.144	[3347, 3209]	[1813 1534] [1519 1690]
WIN3	1v2v3	RFC10	SCALED	F9	0.539	0.06	[5884, 5939, 5693]	[4424 895 565] [953 2665 2321] [685 2657 2351]
WIN30	1v2v3	RFC10	CALIB	F27	0.537	0.087	[3319, 3347, 3209]	[2689 338 292] [797 1443 1107] [516 1512 1181]
WIN30	2v3	RFC10	SCALED	F9	0.532	0.128	[3347, 3209]	[1914 1433] [1657 1552]
WIN30	2v3	RFC50	RAW	F9	0.529	0.147	[3347, 3209]	[1809 1538] [1540 1669]
WIN30	1v2v3	RFC10	CALIB	F9	0.528	0.079	[3319, 3347, 3209]	[2559 429 331] [843 1338 1166] [597 1311 1301]
WIN60	1v2v3	RFC50	CALIB	F27	0.527	0.123	[95, 96, 91]	[81 11 3] [18 33 45] [13 44 34]
WIN30	2v3	RFC10	SCALED	F27	0.526	0.124	[3347, 3209]	[2048 1299] [1834 1375]
WIN30	2v3	RFC10	RAW	F9	0.526	0.144	[3347, 3209]	[1883 1464] [1636 1573]
WIN60	2v3	RFC10	SCALED	F9	0.525	0.2	[96, 91]	[56 40] [51 40]
WIN60	1v2v3	RFC10	CALIB	F27	0.52	0.152	[95, 96, 91]	[74 10 11] [20 35 41] [14 40 37]
WIN30	1v2v3	RFC50	CALIB	F9	0.52	0.081	[3319, 3347, 3209]	[2572 408 339] [814 1142 1391] [540 1255 1414]
WIN5	2v3	RFC50	RAW	F9	0.516	0.057	[5747, 5509]	[3028 2719] [2724 2785]
WIN30	2v3	RFC10	CALIB	F9	0.515	0.086	[3347, 3209]	[1950 1397] [1796 1413]
WIN10	2v3	RFC10	SCALED	F27	0.515	0.085	[5267, 5049]	[3089 2178] [2843 2206]
WIN60	2v3	RFC50	SCALED	F27	0.515	0.205	[96, 91]	[54 42] [50 41]
WIN30	2v3	RFC50	CALIB	F27	0.514	0.126	[3347, 3209]	[1829 1518] [1619 1590]
WIN3	2v3	RFC50	SCALED	F9	0.513	0.053	[5939, 5693]	[3241 2698] [2956 2737]
WIN5	2v3	RFC10	RAW	F9	0.512	0.052	[5747, 5509]	[3227 2520] [2972 2537]
WIN10	2v3	RFC50	RAW	F27	0.511	0.094	[5267, 5049]	[2995 2272] [2744 2305]
WIN10	1v2v3	RFC50	CALIB	F27	0.51	0.057	[5219, 5267, 5049]	[4000 582 637] [1206 1745 2316] [880 2003 2166]
WIN5	2v3	RFC50	SCALED	F27	0.51	0.082	[5747, 5509]	[3080 2667] [2843 2666]
WIN60	2v3	RFC10	SCALED	F27	0.51	0.229	[96, 91]	[57 39] [54 37]
WIN3	2v3	RFC10	SCALED	F9	0.51	0.045	[5939, 5693]	[3493 2446] [3253 2440]
WIN10	2v3	RFC10	RAW	F9	0.51	0.069	[5267, 5049]	[2964 2303] [2725 2324]
WIN3	2v3	RFC10	SCALED	F27	0.509	0.059	[5939, 5693]	[3564 2375] [3340 2353]
WIN60	2v3	RFC50	SCALED	F9	0.509	0.186	[96, 91]	[51 45] [48 43]
WIN10	2v3	RFC50	RAW	F9	0.508	0.081	[5267, 5049]	[2794 2473] [2563 2486]
WIN10	1v2v3	RFC10	CALIB	F27	0.508	0.057	[5219, 5267, 5049]	[3976 653 590] [1228 2009 2030] [966 2192 1891]
WIN10	2v3	RFC50	SCALED	F27	0.508	0.097	[5267, 5049]	[2862 2405] [2678 2371]
WIN5	2v3	RFC50	RAW	F27	0.508	0.066	[5747, 5509]	[3169 2578] [2932 2577]
WIN5	2v3	RFC10	RAW	F27	0.507	0.062	[5747, 5509]	[3371 2376] [3176 2333]
WIN3	2v3	RFC50	SCALED	F27	0.507	0.074	[5939, 5693]	[3215 2724] [2997 2696]
WIN10	2v3	RFC10	RAW	F27	0.507	0.077	[5267, 5049]	[3188 2079] [3004 2045]
WIN5	2v3	RFC10	SCALED	F27	0.505	0.067	[5747, 5509]	[3369 2378] [3208 2301]
WIN30	2v3	RFC50	CALIB	F9	0.504	0.099	[3347, 3209]	[1679 1668] [1575 1634]

Table A32. Cross validation results: all aproaches 2

win size	classes	model	scaling approach	feature approach	acc mean	acc sd	class sizes	confusion matrix
WIN3	2v3	RFC10	RAW	F27	0.504	0.057	[5939, 5693]	[3595 2344] [3427 2266]
WIN60	1v2v3	RFC10	RAW	F27	0.503	0.142	[95, 96, 91]	[67 14 14] [23 43 30] [19 41 31]
WIN5	2v3	RFC10	SCALED	F9	0.503	0.058	[5747, 5509]	[3352 2395] [3205 2304]
WIN5	1v2v3	RFC50	CALIB	F27	0.503	0.058	[5694, 5747, 5509]	[4245 747 702] [1350 1986 2411] [1006 2220 2283]
WIN60	2v3	RFC50	RAW	F9	0.503	0.185	[96, 91]	[50 46] [48 43]
WIN10	2v3	RFC10	CALIB	F27	0.502	0.057	[5267, 5049]	[2909 2358] [2782 2267]
WIN3	2v3	RFC50	RAW	F27	0.502	0.062	[5939, 5693]	[3306 2633] [3153 2540]
WIN30	2v3	RFC10	CALIB	F27	0.501	0.112	[3347, 3209]	[1862 1485] [1748 1461]
WIN10	2v3	RFC50	CALIB	F27	0.501	0.066	[5267, 5049]	[2658 2609] [2517 2532]
WIN5	2v3	RFC50	SCALED	F9	0.5	0.063	[5747, 5509]	[3043 2704] [2906 2603]
WIN60	2v3	RFC10	CALIB	F27	0.499	0.163	[96, 91]	[50 46] [49 42]
WIN5	2v3	RFC10	CALIB	F27	0.498	0.054	[5747, 5509]	[3271 2476] [3184 2325]
WIN3	2v3	RFC10	CALIB	F27	0.497	0.057	[5939, 5693]	[3453 2486] [3372 2321]
WIN3	2v3	RFC50	CALIB	F27	0.497	0.068	[5939, 5693]	[3080 2859] [2965 2728]
WIN3	2v3	RFC10	RAW	F9	0.497	0.054	[5939, 5693]	[3328 2611] [3238 2455]
WIN60	1v2v3	RFC10	CALIB	F9	0.495	0.107	[95, 96, 91]	[75 14 6] [22 32 42] [10 48 33]
WIN5	1v2v3	RFC10	CALIB	F27	0.495	0.061	[5694, 5747, 5509]	[4172 846 676] [1453 2170 2124] [1106 2373 2030]
WIN3	1v2v3	RFC50	CALIB	F27	0.494	0.056	[5884, 5939, 5693]	[4292 871 721] [1452 1994 2493] [1059 2258 2376]
WIN10	2v3	RFC10	SCALED	F9	0.494	0.059	[5267, 5049]	[2906 2361] [2871 2178]
WIN3	2v3	RFC50	RAW	F9	0.493	0.059	[5939, 5693]	[3072 2867] [3008 2685]
WIN60	2v3	RFC10	RAW	F9	0.493	0.185	[96, 91]	[55 41] [55 36]
WIN30	2v3	RFC50	RAW	F27	0.493	0.134	[3347, 3209]	[1806 1541] [1790 1419]
WIN5	2v3	RFC50	CALIB	F9	0.492	0.047	[5747, 5509]	[2927 2820] [2881 2628]
WIN3	2v3	RFC10	CALIB	F9	0.491	0.047	[5939, 5693]	[3330 2609] [3305 2388]
WIN10	2v3	RFC10	CALIB	F9	0.491	0.064	[5267, 5049]	[2823 2444] [2810 2239]
WIN10	1v2v3	RFC10	CALIB	F9	0.49	0.054	[5219, 5267, 5049]	[3728 798 693] [1364 1931 1972] [1036 2080 1933]
WIN10	1v2v3	RFC50	CALIB	F9	0.49	0.055	[5219, 5267, 5049]	[3719 810 690] [1340 1771 2156] [956 1989 2104]
WIN10	2v3	RFC50	CALIB	F9	0.488	0.069	[5267, 5049]	[2592 2675] [2594 2455]
WIN5	2v3	RFC10	CALIB	F9	0.488	0.045	[5747, 5509]	[3216 2531] [3223 2286]
WIN3	1v2v3	RFC10	CALIB	F27	0.486	0.053	[5884, 5939, 5693]	[4165 1002 717] [1506 2247 2186] [1136 2461 2096]
WIN5	2v3	RFC50	CALIB	F27	0.486	0.075	[5747, 5509]	[2869 2878] [2893 2616]
WIN60	1v2v3	RFC50	CALIB	F9	0.485	0.133	[95, 96, 91]	[73 14 8] [23 29 44] [12 44 35]
WIN5	1v2v3	RFC50	CALIB	F9	0.485	0.046	[5694, 5747, 5509]	[3984 913 797] [1502 2002 2243] [1067 2204 2238]
WIN3	2v3	RFC50	CALIB	F9	0.484	0.058	[5939, 5693]	[3005 2934] [3045 2648]
WIN3	1v2v3	RFC50	CALIB	F9	0.482	0.047	[5884, 5939, 5693]	[4075 1004 805] [1526 2045 2368] [1087 2286 2320]
WIN10	2v3	RFC50	SCALED	F9	0.482	0.07	[5267, 5049]	[2584 2683] [2659 2390]
WIN5	1v2v3	RFC10	CALIB	F9	0.481	0.041	[5694, 5747, 5509]	[4002 935 757] [1500 2145 2102] [1133 2386 1990]
WIN3	1v2v3	RFC10	CALIB	F9	0.479	0.05	[5884, 5939, 5693]	[4009 1102 773] [1570 2265 2104] [1158 2423 2112]
WIN60	2v3	RFC50	CALIB	F9	0.475	0.171	[96, 91]	[48 48] [50 41]
WIN30	2v3	RFC10	RAW	F27	0.473	0.123	[3347, 3209]	[1852 1495] [1983 1226]
WIN60	2v3	RFC10	CALIB	F9	0.465	0.157	[96, 91]	[54 42] [58 33]
WIN30	1v2v3	RFC10	RAW	F9	0.462	0.096	[3319, 3347, 3209]	[2006 660 653] [820 1458 1069] [1012 1106 1091]
WIN30	1v2v3	RFC50	RAW	F9	0.459	0.093	[3319, 3347, 3209]	[1952 655 712] [831 1365 1151] [922 1075 1212]
WIN30	1v2v3	RFC10	RAW	F27	0.458	0.112	[3319, 3347, 3209]	[2133 578 608] [1020 1410 917] [964 1286 959]
WIN60	2v3	RFC50	CALIB	F27	0.456	0.123	[96, 91]	[47 49] [53 38]
WIN30	1v2v3	RFC50	RAW	F27	0.454	0.115	[3319, 3347, 3209]	[2255 524 540] [1031 1245 1071] [923 1321 965]
WIN10	1v2v3	RFC50	RAW	F27	0.454	0.076	[5219, 5267, 5049]	[3288 1040 891] [1464 2012 1791] [1393 1918 1738]
WIN5	1v2v3	RFC50	RAW	F27	0.447	0.067	[5694, 5747, 5509]	[3500 1194 1000] [1599 2075 2073] [1435 2086 1988]
WIN10	1v2v3	RFC10	RAW	F27	0.442	0.073	[5219, 5267, 5049]	[3249 1070 900] [1514 2030 1723] [1530 1946 1573]
WIN3	1v2v3	RFC50	RAW	F27	0.44	0.06	[5884, 5939, 5693]	[3546 1288 1050] [1624 2154 2161] [1541 2162 1990]
WIN5	1v2v3	RFC10	RAW	F27	0.44	0.072	[5694, 5747, 5509]	[3446 1194 1054] [1650 2160 1937] [1551 2137 1821]
WIN5	1v2v3	RFC50	RAW	F9	0.435	0.062	[5694, 5747, 5509]	[3170 1383 1141] [1532 2119 2096] [1460 1978 2071]
WIN3	1v2v3	RFC10	RAW	F27	0.434	0.061	[5884, 5939, 5693]	[3525 1311 1048] [1775 2206 1958] [1661 2159 1873]
WIN10	1v2v3	RFC50	RAW	F9	0.434	0.064	[5219, 5267, 5049]	[2942 1177 1100] [1410 1973 1884] [1436 1788 1825]
WIN10	1v2v3	RFC10	RAW	F9	0.428	0.062	[5219, 5267, 5049]	[2964 1211 1044] [1480 1998 1789] [1554 1796 1699]
WIN5	1v2v3	RFC10	RAW	F9	0.427	0.057	[5694, 5747, 5509]	[3221 1385 1088] [1642 2105 2000] [1560 2049 1900]
WIN60	1v2v3	RFC10	RAW	F9	0.416	0.142	[95, 96, 91]	[57 18 20] [24 35 37] [36 31 24]
WIN60	1v2v3	RFC50	RAW	F9	0.414	0.157	[95, 96, 91]	[56 17 22] [23 35 38] [34 32 25]
WIN3	1v2v3	RFC10	RAW	F9	0.413	0.051	[5884, 5939, 5693]	[3202 1501 1181] [1747 2108 2084] [1663 2122 1908]
WIN3	1v2v3	RFC50	RAW	F9	0.411	0.05	[5884, 5939, 5693]	[3117 1492 1275] [1655 2062 2222] [1576 2107 2010]

Table A33. Cross validation results: all aproaches 3

win size	classes	model	scaling approach	feature approach	acc mean	acc sd	class sizes	confusion matrix
WIN60	1v3	RFC50	SCALED	F27	0.972	0.08	[72, 68]	[71 1] [3 65]
WIN60	1v3	RFC10	SCALED	F27	0.972	0.094	[72, 68]	[72 0] [4 64]
WIN60	1v2	RFC10	SCALED	F9	0.951	0.077	[72, 72]	[69 3] [4 68]
WIN60	1v3	RFC50	SCALED	F9	0.951	0.115	[72, 68]	[68 4] [3 65]
WIN60	1v2	RFC50	SCALED	F27	0.944	0.08	[72, 72]	[68 4] [4 68]
WIN60	1v3	RFC10	SCALED	F9	0.944	0.127	[72, 68]	[68 4] [4 64]
WIN60	1v2	RFC10	SCALED	F27	0.944	0.094	[72, 72]	[68 4] [4 68]
WIN60	1v2	RFC50	SCALED	F9	0.944	0.094	[72, 72]	[69 3] [5 67]
WIN30	1v3	RFC50	SCALED	F27	0.934	0.105	[2515, 2408]	[2401 114] [206 2202]
WIN30	1v3	RFC10	SCALED	F27	0.929	0.101	[2515, 2408]	[2406 109] [240 2168]
WIN30	1v3	RFC10	SCALED	F9	0.924	0.113	[2515, 2408]	[2344 171] [192 2216]
WIN30	1v3	RFC50	SCALED	F9	0.922	0.116	[2515, 2408]	[2327 188] [173 2235]
WIN30	1v2	RFC50	SCALED	F27	0.908	0.101	[2515, 2512]	[2358 157] [304 2208]
WIN10	1v3	RFC50	SCALED	F27	0.905	0.109	[3955, 3788]	[3634 321] [401 3387]
WIN30	1v2	RFC10	SCALED	F27	0.904	0.099	[2515, 2512]	[2380 135] [346 2166]
WIN10	1v3	RFC10	SCALED	F9	0.899	0.109	[3955, 3788]	[3625 330] [448 3340]
WIN10	1v3	RFC50	SCALED	F9	0.898	0.11	[3955, 3788]	[3585 370] [409 3379]
WIN30	1v2	RFC50	SCALED	F9	0.898	0.102	[2515, 2512]	[2311 204] [307 2205]
WIN10	1v3	RFC10	SCALED	F27	0.896	0.111	[3955, 3788]	[3625 330] [464 3324]
WIN30	1v2	RFC10	SCALED	F9	0.896	0.1	[2515, 2512]	[2309 206] [316 2196]
WIN5	1v3	RFC50	SCALED	F27	0.893	0.109	[4315, 4133]	[3901 414] [477 3656]
WIN5	1v3	RFC10	SCALED	F27	0.888	0.109	[4315, 4133]	[3924 391] [551 3582]
WIN3	1v3	RFC50	SCALED	F27	0.885	0.114	[4459, 4271]	[3999 460] [536 3735]
WIN5	1v3	RFC50	SCALED	F9	0.884	0.111	[4315, 4133]	[3856 459] [509 3624]
WIN10	1v2	RFC50	SCALED	F27	0.88	0.095	[3955, 3952]	[3482 473] [472 3480]
WIN5	1v3	RFC10	SCALED	F9	0.88	0.113	[4315, 4133]	[3878 437] [569 3564]
WIN3	1v3	RFC10	SCALED	F27	0.876	0.115	[4459, 4271]	[3999 460] [616 3655]
WIN60	1v3	RFC50	CALIB	F9	0.875	0.123	[72, 68]	[66 6] [11 57]
WIN10	1v2	RFC10	SCALED	F27	0.875	0.096	[3955, 3952]	[3528 427] [559 3393]
WIN3	1v3	RFC50	SCALED	F9	0.874	0.116	[4459, 4271]	[3957 502] [588 3683]
WIN5	1v2	RFC50	SCALED	F27	0.871	0.097	[4315, 4312]	[3773 542] [574 3738]
WIN30	1v3	RFC50	CALIB	F9	0.87	0.115	[2515, 2408]	[2326 189] [447 1961]
WIN10	1v2	RFC50	SCALED	F9	0.869	0.101	[3955, 3952]	[3443 512] [527 3425]
WIN3	1v3	RFC10	SCALED	F9	0.868	0.112	[4459, 4271]	[3977 482] [660 3611]
WIN10	1v2	RFC10	SCALED	F9	0.868	0.099	[3955, 3952]	[3497 458] [583 3369]
WIN5	1v2	RFC50	SCALED	F9	0.861	0.098	[4315, 4312]	[3735 580] [622 3690]
WIN60	1v3	RFC10	CALIB	F27	0.861	0.175	[72, 68]	[64 8] [11 57]
WIN5	1v2	RFC10	SCALED	F27	0.861	0.094	[4315, 4312]	[3792 523] [676 3636]
WIN30	1v3	RFC10	CALIB	F27	0.859	0.134	[2515, 2408]	[2304 211] [473 1935]
WIN3	1v2	RFC50	SCALED	F27	0.859	0.099	[4459, 4456]	[3826 633] [625 3831]
WIN30	1v3	RFC10	CALIB	F9	0.859	0.122	[2515, 2408]	[2310 205] [481 1927]
WIN5	1v2	RFC10	SCALED	F9	0.854	0.097	[4315, 4312]	[3748 567] [694 3618]
WIN60	1v3	RFC10	CALIB	F9	0.854	0.142	[72, 68]	[65 7] [13 55]
WIN60	1v3	RFC50	CALIB	F27	0.854	0.172	[72, 68]	[64 8] [12 56]
WIN30	1v3	RFC50	CALIB	F27	0.852	0.128	[2515, 2408]	[2258 257] [456 1952]
WIN3	1v2	RFC10	SCALED	F27	0.851	0.099	[4459, 4456]	[3857 602] [730 3726]
WIN3	1v2	RFC50	SCALED	F9	0.841	0.102	[4459, 4456]	[3752 707] [706 3750]
WIN3	1v2	RFC10	SCALED	F9	0.838	0.101	[4459, 4456]	[3819 640] [803 3653]
WIN10	1v3	RFC50	CALIB	F27	0.823	0.118	[3955, 3788]	[3416 539] [827 2961]
WIN10	1v3	RFC10	CALIB	F27	0.823	0.121	[3955, 3788]	[3482 473] [901 2887]
WIN30	1v2	RFC50	CALIB	F27	0.814	0.139	[2515, 2512]	[2201 314] [623 1889]
WIN10	1v3	RFC50	CALIB	F9	0.813	0.104	[3955, 3788]	[3380 575] [867 2921]
WIN5	1v3	RFC50	CALIB	F27	0.81	0.105	[4315, 4133]	[3658 657] [937 3196]
WIN10	1v3	RFC10	CALIB	F9	0.806	0.106	[3955, 3788]	[3375 580] [919 2869]
WIN3	1v3	RFC50	CALIB	F27	0.8	0.111	[4459, 4271]	[3723 736] [997 3274]
WIN5	1v3	RFC10	CALIB	F27	0.799	0.105	[4315, 4133]	[3676 639] [1057 3076]
WIN60	1v2	RFC50	CALIB	F27	0.799	0.177	[72, 72]	[60 12] [17 55]
WIN5	1v3	RFC50	CALIB	F9	0.795	0.097	[4315, 4133]	[3591 724] [1008 3125]
WIN60	1v2	RFC10	CALIB	F27	0.792	0.198	[72, 72]	[60 12] [18 54]
WIN30	1v2	RFC10	CALIB	F27	0.791	0.15	[2515, 2512]	[2220 295] [758 1754]

Table A34. Cross validation results without trial run 1: all approaches 0

win size	classes	model	scaling approach	feature approach	acc mean	acc sd	class sizes	confusion matrix
WIN3	1v3	RFC10	CALIB	F27	0.79	0.106	[4459, 4271]	[3755 704] [1126 3145]
WIN30	1v2	RFC50	CALIB	F9	0.788	0.135	[2515, 2512]	[2183 332] [732 1780]
WIN10	1v2	RFC50	CALIB	F27	0.786	0.127	[3955, 3952]	[3355 600] [1091 2861]
WIN3	1v3	RFC50	CALIB	F9	0.784	0.105	[4459, 4271]	[3656 803] [1070 3201]
WIN5	1v3	RFC10	CALIB	F9	0.782	0.097	[4315, 4133]	[3594 721] [1120 3013]
WIN30	1v2	RFC10	CALIB	F9	0.782	0.144	[2515, 2512]	[2188 327] [767 1745]
WIN3	1v3	RFC10	CALIB	F9	0.779	0.102	[4459, 4271]	[3696 763] [1165 3106]
WIN60	1v2	RFC10	CALIB	F9	0.778	0.188	[72, 72]	[58 14] [18 54]
WIN10	1v2	RFC10	CALIB	F27	0.777	0.123	[3955, 3952]	[3408 547] [1215 2737]
WIN60	1v2	RFC50	RAW	F27	0.771	0.176	[72, 72]	[54 18] [15 57]
WIN5	1v2	RFC50	CALIB	F27	0.756	0.13	[4315, 4312]	[3481 834] [1274 3038]
WIN10	1v2	RFC50	CALIB	F9	0.751	0.131	[3955, 3952]	[3258 697] [1269 2683]
WIN60	1v2	RFC50	CALIB	F9	0.75	0.209	[72, 72]	[58 14] [22 50]
WIN3	1v2	RFC50	CALIB	F27	0.744	0.127	[4459, 4456]	[3518 941] [1340 3116]
WIN60	1v3	RFC50	RAW	F27	0.743	0.19	[72, 68]	[56 16] [20 48]
WIN5	1v2	RFC10	CALIB	F27	0.742	0.132	[4315, 4312]	[3530 785] [1439 2873]
WIN10	1v2	RFC10	CALIB	F9	0.742	0.128	[3955, 3952]	[3273 682] [1358 2594]
WIN60	1v2	RFC10	RAW	F27	0.736	0.196	[72, 72]	[56 16] [22 50]
WIN60	1v2	RFC50	RAW	F9	0.736	0.208	[72, 72]	[54 18] [20 52]
WIN5	1v2	RFC50	CALIB	F9	0.736	0.12	[4315, 4312]	[3392 923] [1354 2958]
WIN5	1v2	RFC10	CALIB	F9	0.734	0.12	[4315, 4312]	[3469 846] [1449 2863]
WIN3	1v2	RFC10	CALIB	F27	0.732	0.128	[4459, 4456]	[3552 907] [1485 2971]
WIN3	1v2	RFC50	CALIB	F9	0.732	0.112	[4459, 4456]	[3466 993] [1394 3062]
WIN3	1v2	RFC10	CALIB	F9	0.723	0.114	[4459, 4456]	[3505 954] [1515 2941]
WIN60	1v2	RFC10	RAW	F9	0.722	0.229	[72, 72]	[52 20] [20 52]
WIN60	1v3	RFC10	RAW	F27	0.715	0.2	[72, 68]	[56 16] [24 44]
WIN30	1v3	RFC50	RAW	F27	0.705	0.203	[2515, 2408]	[1890 625] [831 1577]
WIN30	1v3	RFC10	RAW	F27	0.696	0.187	[2515, 2408]	[1898 617] [878 1530]
WIN10	1v3	RFC50	RAW	F27	0.693	0.14	[3955, 3788]	[2954 1001] [1402 2386]
WIN5	1v3	RFC50	RAW	F27	0.689	0.128	[4315, 4133]	[3218 1097] [1552 2581]
WIN10	1v2	RFC50	RAW	F27	0.682	0.154	[3955, 3952]	[2916 1039] [1475 2477]
WIN60	1v3	RFC10	RAW	F9	0.681	0.162	[72, 68]	[54 18] [27 41]
WIN5	1v3	RFC10	RAW	F27	0.679	0.124	[4315, 4133]	[3269 1046] [1690 2443]
WIN10	1v3	RFC10	RAW	F27	0.677	0.133	[3955, 3788]	[3021 934] [1599 2189]
WIN30	1v2v3	RFC50	SCALED	F27	0.675	0.119	[2515, 2512, 2408]	[2338 144 33] [222 1347 943] [141 928 1339]
WIN3	1v3	RFC50	RAW	F27	0.675	0.13	[4459, 4271]	[3259 1200] [1667 2604]
WIN60	1v3	RFC50	RAW	F9	0.674	0.151	[72, 68]	[53 19] [27 41]
WIN5	1v2	RFC10	RAW	F27	0.674	0.144	[4315, 4312]	[3174 1141] [1675 2637]
WIN5	1v2	RFC50	RAW	F27	0.674	0.15	[4315, 4312]	[3120 1195] [1613 2699]
WIN10	1v2	RFC10	RAW	F27	0.672	0.151	[3955, 3952]	[2963 992] [1598 2354]
WIN30	1v2	RFC10	RAW	F9	0.669	0.168	[2515, 2512]	[1790 725] [938 1574]
WIN30	1v2	RFC50	RAW	F9	0.668	0.169	[2515, 2512]	[1710 805] [864 1648]
WIN3	1v3	RFC10	RAW	F27	0.667	0.122	[4459, 4271]	[3334 1125] [1807 2464]
WIN60	1v2v3	RFC10	SCALED	F27	0.664	0.139	[72, 72, 68]	[68 4 0] [4 43 25] [3 36 29]
WIN3	1v2	RFC50	RAW	F27	0.662	0.136	[4459, 4456]	[3109 1350] [1663 2793]
WIN10	1v2	RFC50	RAW	F9	0.661	0.153	[3955, 3952]	[2755 1200] [1477 2475]
WIN30	1v2	RFC50	RAW	F27	0.661	0.169	[2515, 2512]	[1769 746] [959 1553]
WIN30	1v3	RFC10	RAW	F9	0.66	0.162	[2515, 2408]	[1807 708] [989 1419]
WIN5	1v3	RFC50	RAW	F9	0.656	0.106	[4315, 4133]	[3004 1311] [1599 2534]
WIN30	1v2v3	RFC10	SCALED	F27	0.655	0.103	[2515, 2512, 2408]	[2321 161 33] [259 1320 933] [139 1027 1242]
WIN30	1v3	RFC50	RAW	F9	0.654	0.167	[2515, 2408]	[1761 754] [975 1433]
WIN3	1v2	RFC10	RAW	F27	0.653	0.133	[4459, 4456]	[3204 1255] [1834 2622]
WIN5	1v3	RFC10	RAW	F9	0.653	0.101	[4315, 4133]	[3124 1191] [1751 2382]
WIN5	1v2	RFC50	RAW	F9	0.649	0.13	[4315, 4312]	[2914 1401] [1625 2687]
WIN10	1v2	RFC10	RAW	F9	0.649	0.146	[3955, 3952]	[2815 1140] [1632 2320]
WIN5	1v2	RFC10	RAW	F9	0.648	0.122	[4315, 4312]	[3026 1289] [1745 2567]
WIN60	1v2v3	RFC10	SCALED	F9	0.646	0.123	[72, 72, 68]	[69 3 0] [4 40 28] [5 35 28]
WIN10	1v3	RFC50	RAW	F9	0.645	0.114	[3955, 3788]	[2779 1176] [1573 2215]
WIN60	1v2v3	RFC50	SCALED	F27	0.645	0.155	[72, 72, 68]	[68 4 0] [5 40 27] [3 37 28]
WIN10	1v3	RFC10	RAW	F9	0.643	0.117	[3955, 3788]	[2852 1103] [1673 2115]

Table A35. Cross validation results without trial run 1: all approaches 1

win size	classes	model	scaling approach	feature approach	acc mean	acc sd	class sizes	confusion matrix
WIN30	1v2	RFC10	RAW	F27	0.643	0.161	[2515, 2512]	[1857 658] [1138 1374]
WIN3	1v3	RFC50	RAW	F9	0.641	0.111	[4459, 4271]	[3082 1377] [1758 2513]
WIN3	1v2	RFC50	RAW	F9	0.636	0.126	[4459, 4456]	[2920 1539] [1707 2749]
WIN3	1v3	RFC10	RAW	F9	0.634	0.111	[4459, 4271]	[3138 1321] [1872 2399]
WIN30	1v2v3	RFC50	SCALED	F9	0.63	0.112	[2515, 2512, 2408]	[2228 185 102] [210 1322 980] [140 1124 1144]
WIN3	1v2	RFC10	RAW	F9	0.626	0.118	[4459, 4456]	[3017 1442] [1892 2564]
WIN60	1v2v3	RFC50	SCALED	F9	0.618	0.11	[72, 72, 68]	[65 5 2] [4 35 33] [4 33 31]
WIN10	1v2v3	RFC50	SCALED	F27	0.618	0.086	[3955, 3952, 3788]	[3456 408 91] [418 1962 1572] [242 1749 1797]
WIN30	1v2v3	RFC10	SCALED	F9	0.617	0.105	[2515, 2512, 2408]	[2239 166 110] [231 1332 949] [152 1236 1020]
WIN10	1v2v3	RFC10	SCALED	F27	0.611	0.087	[3955, 3952, 3788]	[3414 421 120] [414 2058 1480] [267 1861 1660]
WIN60	2v3	RFC10	RAW	F27	0.603	0.2	[72, 68]	[49 23] [32 36]
WIN5	1v2v3	RFC50	SCALED	F27	0.599	0.086	[4315, 4312, 4133]	[3675 470 170] [511 2000 1801] [340 1840 1953]
WIN5	1v2v3	RFC10	SCALED	F27	0.591	0.077	[4315, 4312, 4133]	[3679 456 180] [562 2122 1628] [365 2047 1721]
WIN10	1v2v3	RFC50	SCALED	F9	0.588	0.084	[3955, 3952, 3788]	[3407 363 185] [451 1743 1758] [289 1770 1729]
WIN3	1v2v3	RFC50	SCALED	F27	0.588	0.078	[4459, 4456, 4271]	[3760 525 174] [556 2057 1843] [358 1980 1933]
WIN10	1v2v3	RFC10	SCALED	F9	0.584	0.082	[3955, 3952, 3788]	[3396 352 207] [463 1867 1622] [317 1911 1560]
WIN5	1v2v3	RFC50	SCALED	F9	0.582	0.081	[4315, 4312, 4133]	[3619 439 257] [544 1934 1834] [341 1924 1868]
WIN5	1v2v3	RFC10	SCALED	F9	0.579	0.075	[4315, 4312, 4133]	[3629 446 240] [568 2074 1670] [363 2094 1676]
WIN3	1v2v3	RFC10	SCALED	F27	0.578	0.069	[4459, 4456, 4271]	[3740 526 193] [609 2137 1710] [379 2154 1738]
WIN3	1v2v3	RFC10	SCALED	F9	0.576	0.073	[4459, 4456, 4271]	[3672 518 269] [646 2159 1651] [414 2100 1757]
WIN3	1v2v3	RFC50	SCALED	F9	0.576	0.073	[4459, 4456, 4271]	[3680 529 250] [611 1984 1861] [385 1952 1934]
WIN30	2v3	RFC50	SCALED	F27	0.57	0.155	[2512, 2408]	[1436 1076] [1043 1365]
WIN30	2v3	RFC10	SCALED	F27	0.567	0.128	[2512, 2408]	[1560 952] [1187 1221]
WIN60	2v3	RFC50	RAW	F27	0.561	0.228	[72, 68]	[43 29] [31 37]
WIN30	1v2v3	RFC10	CALIB	F27	0.561	0.122	[2515, 2512, 2408]	[2073 219 223] [628 949 935] [353 889 1166]
WIN30	1v2v3	RFC50	CALIB	F27	0.558	0.124	[2515, 2512, 2408]	[2114 200 201] [650 922 940] [370 914 1124]
WIN60	1v2v3	RFC50	CALIB	F27	0.557	0.147	[72, 72, 68]	[60 8 4] [17 23 32] [6 27 35]
WIN30	1v2v3	RFC50	CALIB	F9	0.55	0.095	[2515, 2512, 2408]	[2080 269 166] [584 905 1023] [346 951 1111]
WIN60	1v2v3	RFC50	CALIB	F9	0.55	0.157	[72, 72, 68]	[56 13 3] [21 26 25] [6 27 35]
WIN60	1v2v3	RFC50	RAW	F27	0.549	0.172	[72, 72, 68]	[52 8 12] [17 35 20] [18 21 29]
WIN60	2v3	RFC50	SCALED	F9	0.549	0.205	[72, 68]	[41 31] [32 36]
WIN30	2v3	RFC10	SCALED	F9	0.549	0.12	[2512, 2408]	[1515 997] [1219 1189]
WIN10	2v3	RFC50	SCALED	F27	0.548	0.118	[3952, 3788]	[2289 1663] [1864 1924]
WIN60	2v3	RFC10	CALIB	F27	0.547	0.198	[72, 68]	[44 28] [36 32]
WIN60	2v3	RFC50	SCALED	F27	0.547	0.251	[72, 68]	[41 31] [33 35]
WIN30	2v3	RFC50	SCALED	F9	0.544	0.116	[2512, 2408]	[1439 1073] [1162 1246]
WIN10	2v3	RFC10	SCALED	F27	0.54	0.105	[3952, 3788]	[2450 1502] [2080 1708]
WIN10	1v2v3	RFC50	CALIB	F27	0.539	0.089	[3955, 3952, 3788]	[3145 421 389] [903 1445 1604] [669 1410 1709]
WIN30	1v2v3	RFC10	CALIB	F9	0.537	0.078	[2515, 2512, 2408]	[2032 307 176] [600 927 985] [338 1041 1029]
WIN30	2v3	RFC10	CALIB	F27	0.534	0.139	[2512, 2408]	[1528 984] [1290 1118]
WIN30	2v3	RFC50	CALIB	F27	0.531	0.146	[2512, 2408]	[1420 1092] [1186 1222]
WIN10	1v2v3	RFC10	CALIB	F27	0.531	0.08	[3955, 3952, 3788]	[3083 482 390] [977 1542 1433] [712 1495 1581]
WIN10	2v3	RFC50	CALIB	F27	0.527	0.097	[3952, 3788]	[2114 1838] [1801 1987]
WIN10	2v3	RFC10	CALIB	F27	0.523	0.087	[3952, 3788]	[2311 1641] [2042 1746]
WIN5	2v3	RFC50	SCALED	F27	0.523	0.092	[4312, 4133]	[2357 1955] [2062 2071]
WIN5	1v2v3	RFC50	CALIB	F27	0.523	0.083	[4315, 4312, 4133]	[3343 540 432] [1060 1532 1720] [728 1602 1803]
WIN3	1v2v3	RFC50	CALIB	F27	0.522	0.08	[4459, 4456, 4271]	[3413 615 431] [1128 1558 1770] [757 1606 1908]
WIN30	2v3	RFC10	RAW	F9	0.52	0.131	[2512, 2408]	[1469 1043] [1318 1090]
WIN30	2v3	RFC10	CALIB	F9	0.519	0.111	[2512, 2408]	[1369 1143] [1212 1196]
WIN60	2v3	RFC10	SCALED	F27	0.519	0.25	[72, 68]	[41 31] [36 32]
WIN5	2v3	RFC10	RAW	F9	0.515	0.059	[4312, 4133]	[2510 1802] [2298 1835]
WIN5	2v3	RFC10	CALIB	F27	0.515	0.073	[4312, 4133]	[2547 1765] [2334 1799]
WIN5	1v2v3	RFC10	CALIB	F27	0.515	0.085	[4315, 4312, 4133]	[3274 620 421] [1094 1684 1534] [785 1734 1614]
WIN60	2v3	RFC10	SCALED	F9	0.514	0.24	[72, 68]	[42 30] [39 29]
WIN3	1v2v3	RFC10	CALIB	F27	0.514	0.077	[4459, 4456, 4271]	[3308 728 423] [1197 1718 1541] [803 1727 1741]
WIN3	2v3	RFC10	SCALED	F9	0.514	0.059	[4456, 4271]	[2642 1814] [2433 1838]
WIN10	1v2v3	RFC50	CALIB	F9	0.513	0.066	[3955, 3952, 3788]	[2990 531 434] [1049 1309 1594] [717 1382 1689]
WIN10	1v2v3	RFC10	CALIB	F9	0.513	0.07	[3955, 3952, 3788]	[2998 546 411] [1065 1431 1456] [709 1522 1557]
WIN3	2v3	RFC10	SCALED	F27	0.511	0.075	[4456, 4271]	[2672 1784] [2490 1781]
WIN60	2v3	RFC50	CALIB	F27	0.511	0.2	[72, 68]	[38 34] [33 35]

Table A36. Cross validation results without trial run 1: all approaches 2

win size	classes	model	scaling approach	feature approach	acc mean	acc sd	class sizes	confusion matrix
WIN5	2v3	RFC10	SCALED	F27	0.509	0.082	[4312, 4133]	[2547 1765] [2382 1751]
WIN5	1v2v3	RFC50	CALIB	F9	0.507	0.063	[4315, 4312, 4133]	[3192 636 487] [1104 1515 1693] [775 1597 1761]
WIN10	2v3	RFC10	CALIB	F9	0.507	0.065	[3952, 3788]	[2155 1797] [2028 1760]
WIN3	2v3	RFC10	CALIB	F27	0.507	0.073	[4456, 4271]	[2638 1818] [2472 1799]
WIN3	2v3	RFC50	SCALED	F9	0.506	0.069	[4456, 4271]	[2385 2071] [2243 2028]
WIN60	1v2v3	RFC10	CALIB	F27	0.505	0.151	[72, 72, 68]	[57 9 6] [17 25 30] [8 35 25]
WIN3	2v3	RFC50	SCALED	F27	0.505	0.097	[4456, 4271]	[2401 2055] [2260 2011]
WIN5	2v3	RFC50	RAW	F9	0.505	0.067	[4312, 4133]	[2300 2012] [2159 1974]
WIN5	2v3	RFC10	CALIB	F9	0.504	0.062	[4312, 4133]	[2430 1882] [2307 1826]
WIN30	2v3	RFC50	CALIB	F9	0.504	0.127	[2512, 2408]	[1238 1274] [1151 1257]
WIN5	2v3	RFC50	CALIB	F27	0.504	0.092	[4312, 4133]	[2262 2050] [2123 2010]
WIN3	2v3	RFC50	CALIB	F27	0.504	0.091	[4456, 4271]	[2375 2081] [2231 2040]
WIN5	2v3	RFC50	CALIB	F9	0.503	0.068	[4312, 4133]	[2243 2069] [2116 2017]
WIN10	2v3	RFC50	CALIB	F9	0.503	0.071	[3952, 3788]	[1973 1979] [1865 1923]
WIN10	2v3	RFC50	SCALED	F9	0.503	0.103	[3952, 3788]	[2035 1917] [1911 1877]
WIN10	2v3	RFC10	SCALED	F9	0.503	0.085	[3952, 3788]	[2173 1779] [2072 1716]
WIN5	2v3	RFC10	SCALED	F9	0.503	0.058	[4312, 4133]	[2509 1803] [2395 1738]
WIN5	2v3	RFC10	RAW	F27	0.503	0.069	[4312, 4133]	[2538 1774] [2415 1718]
WIN10	2v3	RFC50	RAW	F27	0.503	0.097	[3952, 3788]	[2219 1733] [2089 1699]
WIN5	2v3	RFC50	RAW	F27	0.502	0.067	[4312, 4133]	[2325 1987] [2213 1920]
WIN5	2v3	RFC50	SCALED	F9	0.502	0.068	[4312, 4133]	[2287 2025] [2172 1961]
WIN10	2v3	RFC10	RAW	F27	0.501	0.081	[3952, 3788]	[2351 1601] [2251 1537]
WIN60	1v2v3	RFC10	RAW	F27	0.5	0.164	[72, 72, 68]	[48 13 11] [14 33 25] [15 28 25]
WIN30	2v3	RFC50	RAW	F9	0.498	0.161	[2512, 2408]	[1301 1211] [1245 1163]
WIN60	1v2v3	RFC10	CALIB	F9	0.497	0.175	[72, 72, 68]	[55 14 3] [19 25 28] [11 31 26]
WIN3	2v3	RFC10	CALIB	F9	0.497	0.057	[4456, 4271]	[2492 1964] [2435 1836]
WIN5	1v2v3	RFC10	CALIB	F9	0.497	0.06	[4315, 4312, 4133]	[3149 681 485] [1133 1601 1578] [818 1728 1587]
WIN3	1v2v3	RFC50	CALIB	F9	0.496	0.06	[4459, 4456, 4271]	[3229 708 522] [1136 1514 1806] [788 1686 1797]
WIN3	2v3	RFC50	CALIB	F9	0.496	0.069	[4456, 4271]	[2275 2181] [2215 2056]
WIN10	2v3	RFC50	RAW	F9	0.496	0.077	[3952, 3788]	[2075 1877] [2006 1782]
WIN3	2v3	RFC10	RAW	F27	0.496	0.054	[4456, 4271]	[2627 1829] [2585 1686]
WIN3	1v2v3	RFC10	CALIB	F9	0.495	0.063	[4459, 4456, 4271]	[3198 736 525] [1169 1657 1630] [818 1794 1659]
WIN60	2v3	RFC10	RAW	F9	0.492	0.203	[72, 68]	[41 31] [41 27]
WIN3	2v3	RFC50	RAW	F9	0.492	0.066	[4456, 4271]	[2321 2135] [2286 1985]
WIN10	2v3	RFC10	RAW	F9	0.49	0.072	[3952, 3788]	[2171 1781] [2155 1633]
WIN3	2v3	RFC50	RAW	F27	0.49	0.057	[4456, 4271]	[2426 2030] [2421 1850]
WIN3	2v3	RFC10	RAW	F9	0.49	0.062	[4456, 4271]	[2495 1961] [2486 1785]
WIN30	2v3	RFC50	RAW	F27	0.479	0.144	[2512, 2408]	[1357 1155] [1375 1033]
WIN30	2v3	RFC10	RAW	F27	0.478	0.123	[2512, 2408]	[1410 1102] [1442 966]
WIN60	2v3	RFC50	CALIB	F9	0.465	0.196	[72, 68]	[34 38] [36 32]
WIN60	2v3	RFC10	CALIB	F9	0.46	0.17	[72, 68]	[39 33] [42 26]
WIN60	2v3	RFC50	RAW	F9	0.457	0.219	[72, 68]	[37 35] [40 28]
WIN10	1v2v3	RFC50	RAW	F27	0.452	0.071	[3955, 3952, 3788]	[2593 781 581] [1187 1432 1333] [1148 1403 1237]
WIN5	1v2v3	RFC50	RAW	F27	0.451	0.076	[4315, 4312, 4133]	[2789 873 653] [1305 1496 1511] [1132 1548 1453]
WIN5	1v2v3	RFC10	RAW	F27	0.449	0.079	[4315, 4312, 4133]	[2761 909 645] [1348 1589 1375] [1158 1612 1363]
WIN30	1v2v3	RFC50	RAW	F9	0.448	0.098	[2515, 2512, 2408]	[1504 493 518] [620 1027 865] [713 899 796]
WIN60	1v2v3	RFC10	RAW	F9	0.446	0.152	[72, 72, 68]	[43 12 17] [14 35 23] [25 27 16]
WIN30	1v2v3	RFC50	RAW	F27	0.445	0.119	[2515, 2512, 2408]	[1680 448 387] [848 748 916] [705 826 877]
WIN30	1v2v3	RFC10	RAW	F27	0.445	0.113	[2515, 2512, 2408]	[1628 515 372] [909 784 819] [687 831 890]
WIN3	1v2v3	RFC50	RAW	F27	0.444	0.07	[4459, 4456, 4271]	[2795 981 683] [1272 1568 1616] [1161 1643 1467]
WIN10	1v2v3	RFC10	RAW	F27	0.443	0.066	[3955, 3952, 3788]	[2527 858 570] [1204 1477 1271] [1160 1464 1164]
WIN3	1v2v3	RFC10	RAW	F27	0.442	0.073	[4459, 4456, 4271]	[2779 990 690] [1330 1647 1479] [1228 1665 1378]
WIN5	1v2v3	RFC50	RAW	F9	0.434	0.069	[4315, 4312, 4133]	[2481 1025 809] [1184 1570 1558] [1088 1566 1479]
WIN5	1v2v3	RFC10	RAW	F9	0.433	0.065	[4315, 4312, 4133]	[2507 1069 739] [1266 1605 1441] [1106 1619 1408]
WIN30	1v2v3	RFC10	RAW	F9	0.432	0.098	[2515, 2512, 2408]	[1517 502 496] [642 1007 863] [766 966 676]
WIN10	1v2v3	RFC10	RAW	F9	0.43	0.059	[3955, 3952, 3788]	[2367 901 687] [1077 1495 1380] [1140 1483 1165]
WIN10	1v2v3	RFC50	RAW	F9	0.427	0.057	[3955, 3952, 3788]	[2304 909 742] [1046 1471 1435] [1115 1458 1215]
WIN3	1v2v3	RFC50	RAW	F9	0.422	0.065	[4459, 4456, 4271]	[2491 1122 846] [1283 1579 1594] [1148 1634 1489]
WIN3	1v2v3	RFC10	RAW	F9	0.421	0.061	[4459, 4456, 4271]	[2527 1123 809] [1367 1607 1482] [1186 1674 1411]
WIN60	1v2v3	RFC50	RAW	F9	0.413	0.15	[72, 72, 68]	[42 14 16] [13 29 30] [21 31 16]

Table A37. Cross validation results without trial run 1: all approaches 3

E.2 Individual Models

Participant	1v2v3	1v2	1v3	2v3	1v2v3 CV	1v2 CV	1v3 CV	2v3 CV
P1	0.562	0.723	0.801	0.579	0.462	0.635	0.765	0.527
P2	0.721	0.838	0.877	0.721	0.489	0.758	0.771	0.475
P**3**	nan	0.795	nan	nan	0.486	0.71	0.742	0.498
P4	0.647	0.848	0.859	0.641	0.525	0.84	0.668	0.513
P5	0.764	0.891	0.941	0.744	0.415	0.655	0.586	0.516
P6	0.737	0.887	0.962	0.717	0.487	0.754	0.824	0.488
P8	0.624	0.877	0.84	0.629	0.51	0.866	0.786	0.418
P10	0.559	**0.793**	0.804	0.503	0.53	**0.878**	0.753	0.499
P12	0.536	**0.756**	0.775	0.507	0.546	**0.801**	0.731	0.519
P13	0.65	0.788	0.823	0.675	0.479	0.787	0.801	0.449
P15	**0.788**	0.886	0.915	**0.836**	0.613	0.823	0.901	0.576
P16	0.544	**0.619**	0.688	0.691	0.514	**0.785**	0.859	0.452
P17	0.596	0.833	0.878	0.534	0.459	0.669	0.733	0.508
P**19**	0.535	**0.653**	0.807	0.547	0.48	**0.673**	0.673	0.538
P20	0.556	0.887	0.943	0.401	0.528	0.728	0.813	0.57
P21	0.636	0.941	0.85	0.596	0.516	0.789	0.855	0.454
P22	0.561	0.712	0.885	0.638	0.451	0.626	0.657	0.543
P23	0.483	**0.662**	0.681	0.522	0.522	**0.761**	0.766	0.515
P31	0.592	0.713	0.814	0.668	0.55	0.867	0.895	0.455
P33	0.615	0.884	0.713	0.701	0.562	0.85	0.853	0.511
P35	0.607	0.738	0.871	0.68	0.556	0.657	0.836	0.608
P36	0.446	0.707	0.607	0.548	0.417	0.709	0.639	0.426
P39	0.711	0.844	0.868	0.746	0.493	0.709	0.771	0.487
P48	0.736	0.944	0.871	0.74	0.556	0.718	0.824	0.595
Mean	**0.618**	**0.801**	**0.829**	**0.633**	**0.506**	**0.752**	**0.771**	**0.506**
SD	0.089	0.091	0.087	0.101	0.046	0.076	0.081	0.049
Min	0.446	0.619	0.607	0.401	0.415	0.626	0.586	0.418
Max	0.788	0.944	0.962	0.836	0.613	0.878	0.901	0.608

Table A38. Descriptive statistics specificity individual models and the average validation score when the participant was left out during LOUOCV (indicated by CV).

Participant	1v2v3	1v2	1v3	2v3	1v2v3 CV	1v2 CV	1v3 CV	2v3 CV
1	0.651	0.829	0.923	0.651	0.517	0.672	0.84	0.527
2	0.671	0.829	0.816	0.668	0.502	0.802	0.764	0.468
3	nan	0.824	nan	nan	0.492	0.695	0.755	0.496
4	0.674	0.897	0.92	0.617	0.53	0.855	0.713	0.494
5	0.761	0.974	0.981	0.628	0.473	0.719	0.654	0.545
6	0.712	0.898	0.954	0.686	0.516	0.765	0.815	0.508
8	0.7	0.886	0.883	0.655	0.5	0.87	0.78	0.405
10	0.546	0.79	0.765	0.522	0.506	0.82	0.764	0.477
12	0.569	0.88	0.858	0.484	0.548	0.854	0.788	0.447
13	0.748	0.889	0.989	0.718	0.57	0.836	0.882	0.472
15	0.884	0.936	0.981	0.903	0.626	0.841	0.916	0.581
16	0.489	0.613	0.636	0.638	0.513	0.816	0.856	0.448
17	0.705	0.939	0.976	0.578	0.49	0.675	0.743	0.543
19	0.647	0.819	0.949	0.533	0.46	0.649	0.671	0.575
20	0.491	0.894	0.943	0.297	0.578	0.751	0.845	0.564
21	0.668	0.996	0.97	0.539	0.568	0.812	0.922	0.5
22	0.555	0.629	0.821	0.707	0.472	0.668	0.691	0.555
23	0.506	0.773	0.778	0.436	0.572	0.833	0.811	0.508
31	0.585	0.74	0.87	0.584	0.578	0.894	0.916	0.498
33	0.696	0.935	0.801	0.75	0.564	0.875	0.909	0.498
35	0.72	0.733	0.962	0.774	0.532	0.623	0.824	0.623
36	0.431	0.744	0.617	0.507	0.443	0.766	0.736	0.441
39	0.785	0.837	0.851	0.84	0.534	0.761	0.801	0.514
48	0.766	0.929	0.834	0.81	0.565	0.734	0.871	0.616
Mean	0.65	0.842	0.873	0.632	0.527	0.774	0.803	0.513
SD	0.11	0.098	0.103	0.136	0.044	0.078	0.077	0.054
Min	0.431	0.613	0.617	0.297	0.443	0.623	0.654	0.405
Max	0.884	0.996	0.989	0.903	0.626	0.894	0.922	0.623

Table A39. Descriptive statistics specificity individual models (without trial run 1) and the average validation score when the participant was left out during LOUOCV (indicated by CV).

E.3 Descriptives and Inferential Statistics

E.3.1 Comparing Window Sizes

win size	1v2v3	1v2	1v3	2v3
WIN60	0.524	0.81	0.814	0.508
WIN30	0.528	0.758	0.786	0.517
WIN10	0.499	0.742	0.76	0.502
WIN5	0.494	0.731	0.753	0.502
WIN3	0.485	0.719	0.742	0.5
All				
Count	60	60	60	60
SD	0.057	0.074	0.085	0.021
Min	0.411	0.618	0.613	0.456
Max	0.62	0.891	0.901	0.583
Q25	0.457	0.685	0.678	0.494
Q50 (Median)	0.503	0.75	0.795	0.505
Q75	0.55	0.814	0.839	0.512
WIN60				
Count	12	12	12	12
SD	0.066	0.064	0.077	0.036
Min	0.414	0.693	0.67	0.456
Max	0.613	0.891	0.901	0.583
Q25	0.493	0.785	0.782	0.488
Q50 (Median)	0.524	0.802	0.842	0.506
Q75	0.589	0.86	0.871	0.517
WIN30				
Count	12	12	12	12
SD	0.059	0.07	0.085	0.022
Min	0.454	0.667	0.649	0.473
Max	0.62	0.86	0.867	0.556
Q25	0.461	0.685	0.701	0.503
Q50 (Median)	0.532	0.761	0.831	0.521
Q75	0.574	0.815	0.841	0.53
WIN10				
Count	12	12	12	12
SD	0.051	0.071	0.089	0.011
Min	0.428	0.646	0.623	0.482
Max	0.571	0.838	0.857	0.515
Q25	0.451	0.677	0.665	0.493
Q50 (Median)	0.499	0.736	0.79	0.505
Q75	0.55	0.812	0.834	0.509
WIN5				
Count	12	12	12	12
SD	0.05	0.071	0.08	0.01
Min	0.427	0.632	0.635	0.486
Max	0.562	0.826	0.844	0.516
Q25	0.445	0.67	0.667	0.497
Q50 (Median)	0.49	0.724	0.775	0.504
Q75	0.548	0.806	0.823	0.509
WIN3				
Count	12	12	12	12
SD	0.053	0.07	0.086	0.009
Min	0.411	0.618	0.613	0.484
Max	0.555	0.815	0.836	0.513
Q25	0.438	0.659	0.653	0.496
Q50 (Median)	0.484	0.711	0.767	0.499
Q75	0.54	0.794	0.819	0.508

Table A40. Comparing Window Size

classes	T	p
1v2v3	36.333	p < 0.001
1v2	47.267	p < 0.001
1v3	42.867	p < 0.001
2v3	11.067	p = 0.026

Table A41. Comparing Window Size: Friedman

classes	comp1	comp2	W	p	Rank-Biserial
1v2v3	WIN60	WIN30	28.0	p = 0.388	-0.282
1v2v3	WIN60	WIN10	11.0	p = 0.028	0.718
1v2v3	WIN60	WIN5	7.0	p = 0.012	0.821
1v2v3	WIN60	WIN3	0.0	p = 0.002	1.0
1v2v3	WIN30	WIN10	1.0	p = 0.003	0.974
1v2v3	WIN30	WIN5	0.0	p = 0.002	1.0
1v2v3	WIN30	WIN3	0.0	p = 0.002	1.0
1v2v3	WIN10	WIN5	3.0	p = 0.005	0.923
1v2v3	WIN10	WIN3	0.0	p = 0.002	1.0
1v2v3	WIN5	WIN3	0.0	p = 0.002	1.0
1v2	WIN60	WIN30	0.0	p = 0.002	1.0
1v2	WIN60	WIN10	0.0	p = 0.002	1.0
1v2	WIN60	WIN5	0.0	p = 0.002	1.0
1v2	WIN60	WIN3	0.0	p = 0.002	1.0
1v2	WIN30	WIN10	1.0	p = 0.003	0.974
1v2	WIN30	WIN5	0.0	p = 0.002	1.0
1v2	WIN30	WIN3	0.0	p = 0.002	1.0
1v2	WIN10	WIN5	0.0	p = 0.002	1.0
1v2	WIN10	WIN3	0.0	p = 0.002	1.0
1v2	WIN5	WIN3	0.0	p = 0.002	1.0
1v3	WIN60	WIN30	6.0	p = 0.01	0.846
1v3	WIN60	WIN10	0.0	p = 0.002	1.0
1v3	WIN60	WIN5	0.0	p = 0.002	1.0
1v3	WIN60	WIN3	0.0	p = 0.002	1.0
1v3	WIN30	WIN10	3.0	p = 0.005	0.923
1v3	WIN30	WIN5	0.0	p = 0.002	1.0
1v3	WIN30	WIN3	0.0	p = 0.002	1.0
1v3	WIN10	WIN5	13.0	p = 0.041	0.667
1v3	WIN10	WIN3	0.0	p = 0.002	1.0
1v3	WIN5	WIN3	0.0	p = 0.002	1.0
2v3	WIN60	WIN30	23.0	p = 0.209	-0.41
2v3	WIN60	WIN10	34.0	p = 0.695	0.128
2v3	WIN60	WIN5	37.0	p = 0.875	0.051
2v3	WIN60	WIN3	32.0	p = 0.583	0.179
2v3	WIN30	WIN10	16.0	p = 0.071	0.59
2v3	WIN30	WIN5	15.0	p = 0.06	0.615
2v3	WIN30	WIN3	12.0	p = 0.034	0.692
2v3	WIN10	WIN5	37.0	p = 0.875	-0.051
2v3	WIN10	WIN3	24.0	p = 0.239	0.385
2v3	WIN5	WIN3	34.0	p = 0.695	0.128

Table A42. Comparing Window Size: Wilcoxon

E.3.2 Comparing Scaling

scaling approach	1v2v3	1v2	1v3	2v3
RAW	0.446	0.676	0.665	0.512
CALIB	0.503	0.749	0.802	0.492
SCALED	0.569	0.832	0.847	0.513
All				
Count	60	60	60	60
SD	0.057	0.074	0.085	0.021
Min	0.411	0.618	0.613	0.456
Max	0.62	0.891	0.901	0.583
Q25	0.457	0.685	0.678	0.494
Q50 (Median)	0.503	0.75	0.795	0.505
Q75	0.55	0.814	0.839	0.512
RAW				
Count	20	20	20	20
SD	0.031	0.045	0.043	0.024
Min	0.411	0.618	0.613	0.473
Max	0.542	0.791	0.795	0.583
Q25	0.428	0.651	0.638	0.5
Q50 (Median)	0.44	0.665	0.656	0.508
Q75	0.455	0.683	0.673	0.513
CALIB				
Count	20	20	20	20
SD	0.021	0.037	0.032	0.014
Min	0.479	0.707	0.755	0.456
Max	0.548	0.843	0.847	0.515
Q25	0.486	0.721	0.776	0.488
Q50 (Median)	0.495	0.741	0.793	0.495
Q75	0.52	0.772	0.835	0.501
SCALED				
Count	20	20	20	20
SD	0.025	0.029	0.023	0.016
Min	0.539	0.791	0.818	0.482
Max	0.62	0.891	0.901	0.556
Q25	0.55	0.812	0.831	0.507
Q50 (Median)	0.561	0.822	0.841	0.51
Q75	0.589	0.854	0.861	0.518

Table A43. Comparing Scaling

classes	T	p
1v2v3	38.1	$p < 0.001$
1v2	40.0	$p < 0.001$
1v3	40.0	$p < 0.001$
2v3	22.3	$p < 0.001$

Table A44. Comparing Scaling: Friedman

classes	comp1	comp2	W	p	Rank-Biserial
1v2v3	RAW	CALIB	1.0	$p < 0.001$	-0.99
1v2v3	RAW	SCALED	0.0	$p < 0.001$	-1.0
1v2v3	CALIB	SCALED	0.0	$p < 0.001$	-1.0
1v2	RAW	CALIB	0.0	$p < 0.001$	-1.0
1v2	RAW	SCALED	0.0	$p < 0.001$	-1.0
1v2	CALIB	SCALED	0.0	$p < 0.001$	-1.0
1v3	RAW	CALIB	0.0	$p < 0.001$	-1.0
1v3	RAW	SCALED	0.0	$p < 0.001$	-1.0
1v3	CALIB	SCALED	0.0	$p < 0.001$	-1.0
2v3	RAW	CALIB	29.0	$p = 0.005$	0.724
2v3	RAW	SCALED	90.0	$p = 0.575$	-0.143
2v3	CALIB	SCALED	2.0	$p < 0.001$	-0.981

Table A45. Comparing Scaling: Wilcoxon

E.3.3 Comparing Features

feature approach	1v2v3	1v2	1v3	2v3
FEATURES9	0.495	0.738	0.759	0.503
FEATURES27	0.517	0.766	0.783	0.509
All				
Count	60	60	60	60
SD	0.057	0.074	0.085	0.021
Min	0.411	0.618	0.613	0.456
Max	0.62	0.891	0.901	0.583
Q25	0.457	0.685	0.678	0.494
Q50 (Median)	0.503	0.75	0.795	0.505
Q75	0.55	0.814	0.839	0.512
FEATURES9				
Count	30	30	30	30
SD	0.057	0.073	0.09	0.017
Min	0.411	0.618	0.613	0.465
Max	0.589	0.879	0.885	0.54
Q25	0.441	0.688	0.66	0.492
Q50 (Median)	0.487	0.723	0.781	0.503
Q75	0.548	0.803	0.831	0.513
FEATURES27				
Count	30	30	30	30
SD	0.055	0.074	0.079	0.024
Min	0.434	0.652	0.644	0.456
Max	0.62	0.891	0.901	0.583
Q25	0.465	0.689	0.716	0.499
Q50 (Median)	0.515	0.772	0.801	0.507
Q75	0.555	0.824	0.844	0.511

Table A46. Comparing Features

classes	comp1	comp2	W	p	Rank-Biserial
1v2v3	FEATURES9	FEATURES27	6.0	$p < 0.001$	-0.974
1v2	FEATURES9	FEATURES27	2.0	$p < 0.001$	-0.991
1v3	FEATURES9	FEATURES27	18.0	$p < 0.001$	-0.923
2v3	FEATURES9	FEATURES27	166.0	$p = 0.171$	-0.286

Table A47. Comparing Features: Wilcoxon

E.3.4 Comparing Number of Trees

model	1v2v3	1v2	1v3	2v3
RFC10	0.503	0.748	0.766	0.506
RFC50	0.509	0.756	0.776	0.505
All				
Count	60	60	60	60
SD	0.057	0.074	0.085	0.021
Min	0.411	0.618	0.613	0.456
Max	0.62	0.891	0.901	0.583
Q25	0.457	0.685	0.678	0.494
Q50 (Median)	0.503	0.75	0.795	0.505
Q75	0.55	0.814	0.839	0.512
RFC10				
Count	30	30	30	30
SD	0.056	0.076	0.083	0.02
Min	0.413	0.618	0.613	0.465
Max	0.606	0.885	0.875	0.583
Q25	0.459	0.68	0.687	0.497
Q50 (Median)	0.499	0.746	0.787	0.505
Q75	0.546	0.808	0.831	0.512
RFC50				
Count	30	30	30	30
SD	0.059	0.074	0.087	0.022
Min	0.411	0.632	0.617	0.456
Max	0.62	0.891	0.901	0.561
Q25	0.456	0.692	0.673	0.493
Q50 (Median)	0.506	0.755	0.8	0.505
Q75	0.55	0.815	0.842	0.512

Table A48. Comparing number of trees

classes	comp1	comp2	W	p	Rank-Biserial
1v2v3	RFC10	RFC50	80.0	$p = 0.002$	-0.656
1v2	RFC10	RFC50	33.0	$p < 0.001$	-0.858
1v3	RFC10	RFC50	46.0	$p < 0.001$	-0.802
2v3	RFC10	RFC50	229.0	$p = 0.943$	0.015

Table A49. Comparing number of trees: Wilcoxon

E.4 Plots

E.4.1 Comparison Plots

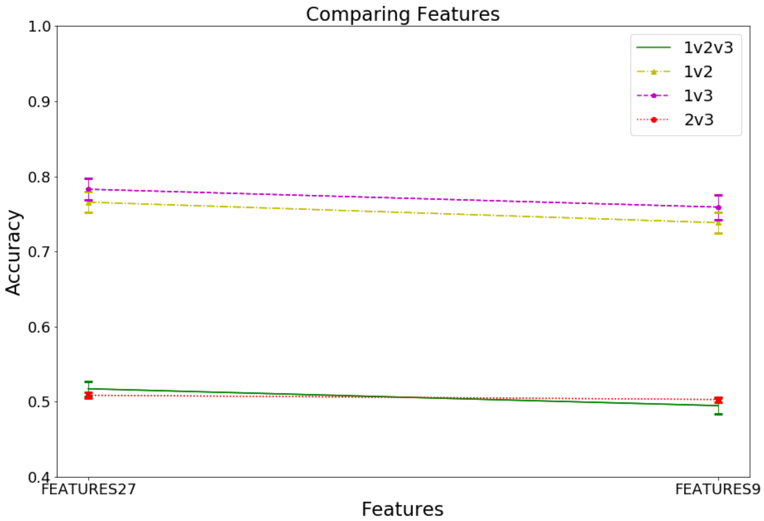


Fig. A1. Features

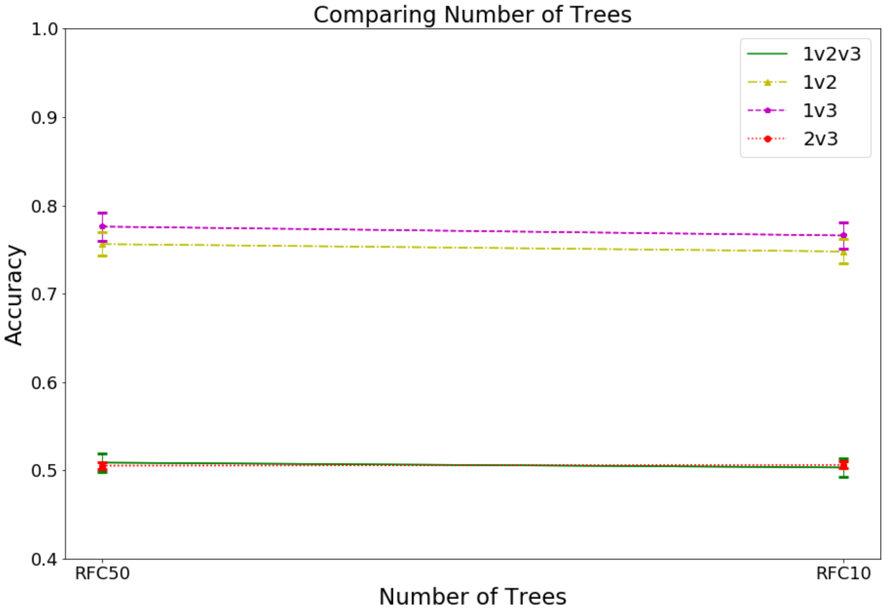


Fig. A2. Number of Trees

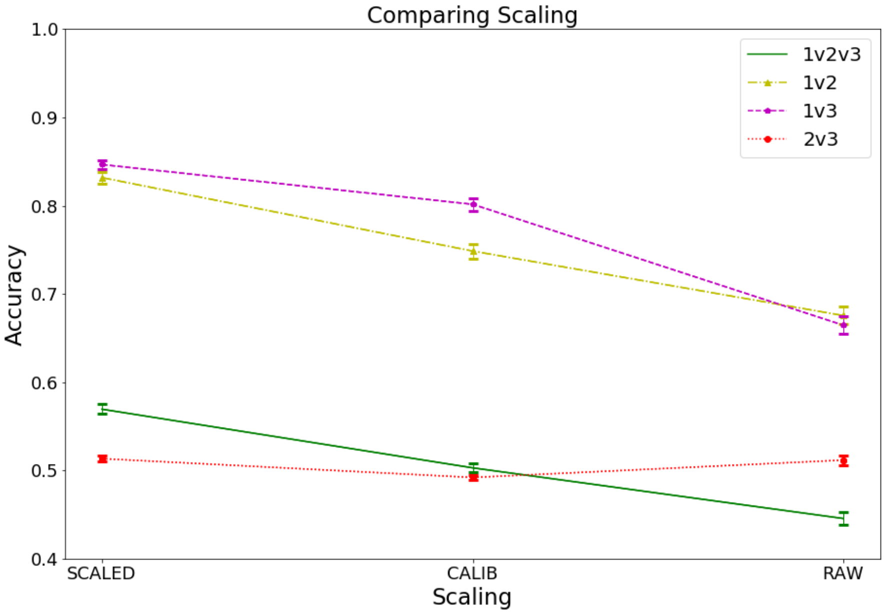


Fig. A3. Scaling

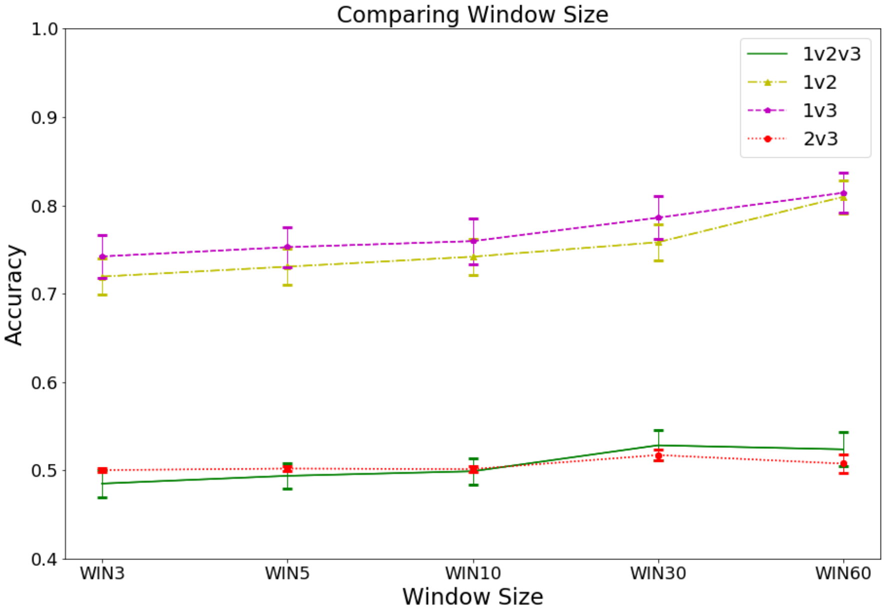


Fig. A4. Window Size

E.4.2 Different Pipelines Window Sizes/Classes Development

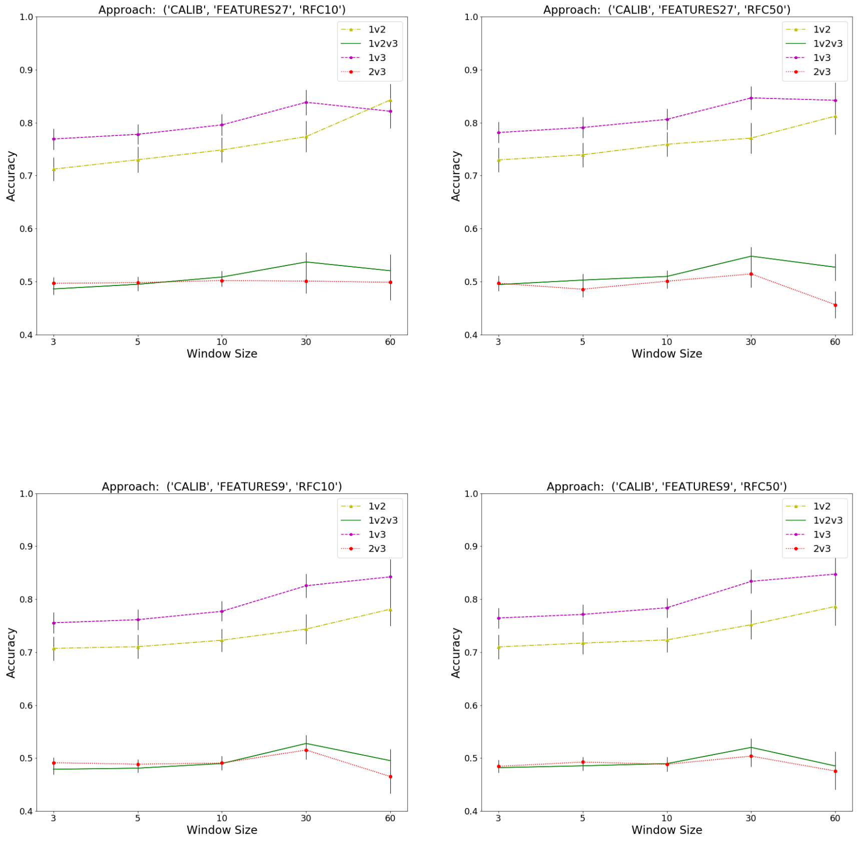


Fig. A5. *CALIB*

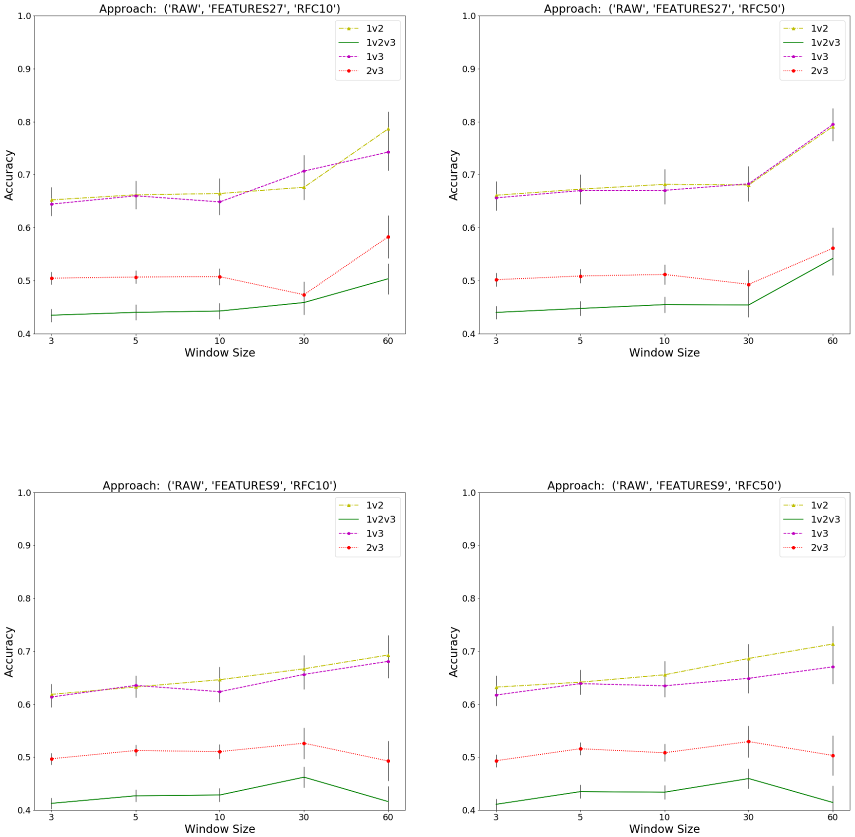

Approach: ('RAW', 'FEATURES27', 'RFC10')
Approach: ('RAW', 'FEATURES27', 'RFC50')
Approach: ('RAW', 'FEATURES9', 'RFC10')
Approach: ('RAW', 'FEATURES9', 'RFC50')
Accuracy
Window Size
1.0
0.9
0.8
0.7
0.6
0.5
0.4
3
5
10
30
60
1v2
1v2v3
1v3
2v3

Fig. A6. *RAW*

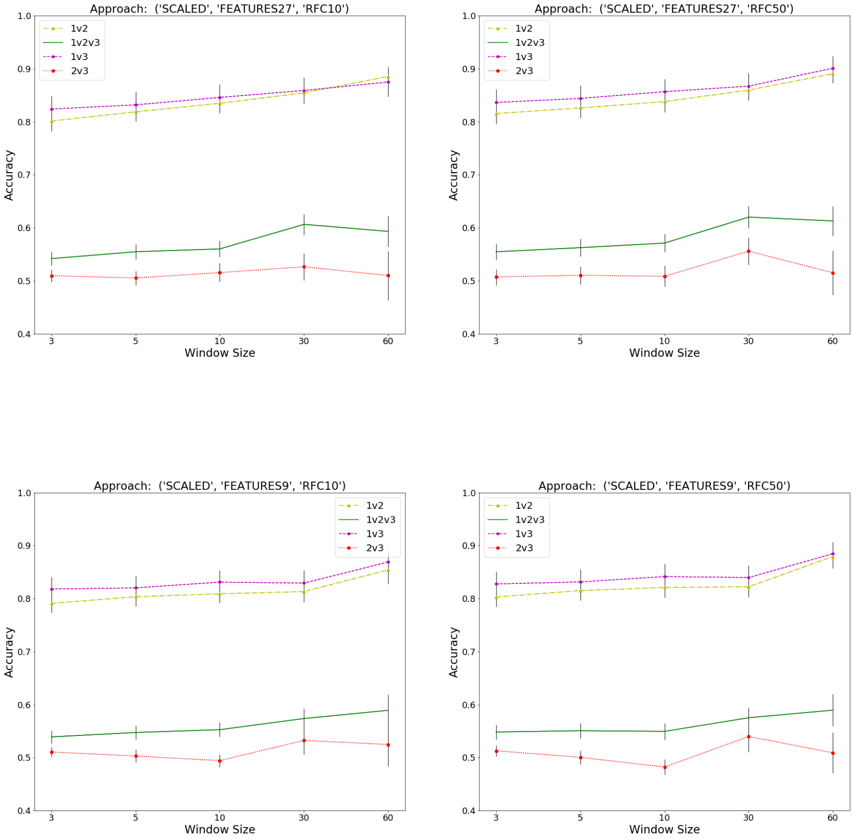


Fig. A7. *SCALED*

F PERFORMANCE STUDY 2

	Accuracy							
	LETTER		IMG		SPATIAL		AUDIO	
	1-back	2-back	1-back	2-back	1-back	2-back	1-back	2-back
49	0.946	0.914	0.968	0.914	0.946	0.903	0.957	0.935
50	1.0	0.903	0.978	0.946	0.989	0.882	1.0	0.914
51	0.968	0.957	0.978	0.925	0.978	0.957	0.946	0.968
52	1.0	0.989	1.0	0.968	0.968	0.946	0.989	0.968
53	0.968	0.957	0.957	0.903	0.935	0.978	0.978	0.978
54	0.978	0.903	0.989	0.903	0.978	0.925	0.957	0.871
Mean	0.977	0.937	0.978	0.927	0.966	0.932	0.971	0.939
Median	0.973	0.935	0.978	0.919	0.973	0.935	0.968	0.952
Std. Deviation	0.021	0.036	0.015	0.026	0.021	0.036	0.021	0.041
Range	0.054	0.086	0.043	0.065	0.054	0.097	0.054	0.108

	Response Time							
	LETTER		IMG		SPATIAL		AUDIO	
	1-back	2-back	1-back	2-back	1-back	2-back	1-back	2-back
49	759.054	848.785	646.452	770.72	647.742	660.581	832.172	815.409
50	610.849	756.333	626.645	633.581	697.731	733.344	489.634	617.043
51	575.602	844.269	695.269	883.699	497.914	629.366	531.333	703.634
52	573.409	701.097	708.946	724.237	540.484	577.505	553.097	591.505
53	630.925	715.925	658.978	791.731	767.054	749.559	638.753	613.806
54	540.075	851.075	505.538	669.882	473.634	719.161	525.409	762.817
Mean	614.986	786.247	640.305	745.642	604.093	678.253	595.066	684.036
Median	593.226	800.301	652.715	747.478	594.113	689.871	542.215	660.339
Std. Deviation	77.354	70.101	72.757	90.076	117.957	67.299	126.427	91.467
Range	218.978	149.978	203.409	250.118	293.419	172.054	342.538	223.903

Table A50. Performance scores (accuracy and response time) for the four n-back task variants.

G CONTENTS OF THE USB FLASH DRIVE

The digital copy of this work can be found on the attached USB flash drive. It includes:

- The raw study data for each data stream.
- The jupyter notebooks with which the data was analysed and machine learning was performed.
- Preprocessed data per participant and results of the trained pipelines.
- JASP files which were used for analysis.
- The E4 Client to collect data from the wristband.
- The SMI Data Collector to collect data from the eye-tracker.
- The Central Data Collector which managed the experiment, and the data collection.
- Lists of python packages used for analysis and the central data collector.
- OpenSesame experiment files for both experiments.
- The study documents for both experiments.
- Author's previous work: seminar and project report on which this thesis was built on.

There exist git repositories for the different components on the university gitlab severs. These can only be cloned using SSH, therefore, a university account is required that needs to be added to the repository (send a request to the author).

- https://git.uni-konstanz.de/philipp-bauer/central_data_collector/
- https://git.uni-konstanz.de/philipp-bauer/data_analysis
- https://git.uni-konstanz.de/philipp-bauer/SMI_Gaze_Collector
- https://git.uni-konstanz.de/philipp-bauer/EmpaticaE4Client