Article

Datasets for Cognitive Load Inference Using Wearable Sensors and Psychological Traits

Martin Gjoreski [1,2,†,*], Tine Kolenik [1,2,†], Timotej Knez [3] Mitja Luštrek [1,2], Matjaž Gams [1,2], Hristijan Gjoreski [4] and Veljko Pejović [3]

1 Jožef Stefan Institute, Slovenia; name.surname@ijs.si
2 Jožef Stefan Postgraduate School, Slovenia
3 University of Ljubljana, Slovenia
4 Ss. Cyril and Methodius University, N. Macedonia
* Correspondence: martin.gjoreski@ijs.si
† These authors contributed equally to this work.

Version May 14, 2020 submitted to Appl. Sci.

Abstract: This study introduces two datasets for multimodal research on cognitive load inference and personality traits. Different to other datasets in Affective Computing, which disregard participants' personality traits, or focus only on emotions, stress, or cognitive load from one specific task, the participants in our experiments performed seven different tasks in total. In the first dataset, 23 participants played a varying difficulty (easy, medium and hard) game on a smartphone. In the second dataset, 23 participants performed six psychological tasks on a PC, again with a varying difficulty. In both experiments, the participants filled personality-traits questionnaire and marked their perceived cognitive load using NASA-TLX after each task. Additionally, the participants' physiological response was recorded using a wrist-device measuring heart rate, beat-to-beat intervals, galvanic skin response, skin temperature, and 3-axis acceleration. The datasets allow multimodal study of physiological responses of individuals in relation to their personality and cognitive load. Various analyses of relationships between personality traits, subjective cognitive load (i.e., NASA-TLX), and objective cognitive load (i.e., task difficult) are presented. Additionally, baseline machine learning models for recognizing task difficulty are presented, including a multi task learning (MTL) neural network that outperforms single task neural network by simultaneously learning from the two datasets. The datasets are publicly available to advance the field of cognitive load inference using commercially available devices.

Keywords: cognitive load; dataset; Affective Computing; machine learning; physiology; personality traits; sensor data.

---

1. Introduction

Affective Computing is the study and development of systems that have the ability to recognize and process human affective states[1]. While sensor-based recognition of the human physical activity has reached a certain level of maturity, e.g., most mobile devices are nowadays capable of counting steps based on acceleration sensors, the human mental state recognition, e.g., stress, mental health and cognitive load, recognition remains challenging. Yet, the demand for advancing affective computing research is rising, since through improved understanding of its human users, affective computing promises to push the frontiers of human-computer interaction (HCI) and to enable new much-needed services that are directly related to psychological states, e.g., mobile healthcare [2]. One of the main impediment factors for the advance of affective computing research is its reliance on real-world user studies that often need to involve elaborate experimental protocols and physiological signal measurements. Being difficult to gather, datasets that include various aspects of human internal states (e.g. the participants' impressions, personality traits, etc.) as well as detailed physiological signal measurements during the experiments are rare and seldom publicly available.

An important aspect of Affective Computing is the cognitive load inference. There are different reasons why ubiquitous computing devices would benefit from being aware of their users' cognitive load, the most important of which is likely to prevent the undesirable effects of attention grabbing at times when a user is occupied with a difficult task. Research has repeatedly shown that improperly timed notifications can be distracting, causing a negative effect on task performance [3–8], increasing stress [9], and reducing well-being [10]. Fortunately, increased cognitive load is reflected in a measurable signal change. When humans experience a psycho-physiological load, e.g., in the form of a demanding task, the activation of the sympathetic nervous system increases [11]. This increased activation translates into changes in the blood pressure [12], heart rate variability [13], respiration [14], brain activity [15], galvanic skin response (GSR) [16,17], eye movement [15], pupil size, facial expressions [18], and other factors. The physiological changes can be measured with special equipment, e.g. a nasal thermistor, a chest respiration strap, ECG (Electrocardiogram), a sphygmomanometer (blood pressure monitor), electroencephalography (EEG), to name a few. Yet, the high cost, bulkiness, and the fact that they work only if a user is static and strapped with sensors, all limit the applicability of these devices in ubiquitous computing. In this paper we focus on two recent data collection campaigns [19,20] that capture the above reactions using off-the-shelf equipment, thus greatly expanding the potential applications in which the knowledge of a user's cognitive load can be harnessed. In terms of the actual equipment, the MS band sensing wristband was used in both studies, as it provides an open API for collecting multimodal data pertaining to Heart rate, RR intervals, GSR, temperature and acceleration.

Personality traits are an important, but often overlooked, mediator of a user's response to increased mental workload. Different psychological profiles, especially those measured with the Big Five Personality Test (aka OCEAN, denoting the five personality dimension: Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) [21], have been shown to respond to cognitive load and its consequences (e.g., stress) differently. The difference manifests in both a user's subjective perception of the workload, as well as in the user's physiological response [22]. Knowing the details about a user's personality traits opens the doors for further technology adaptation. The datasets introduced in this paper include personality traits information of all the participants, in addition to the abovementioned physiological signal samples.

The main contribution of this paper is the introduction of Snake and CogLoad[1], two datasets collected in two separate experiments with 46 participants overall. These datasets are the first that enable multimodal study of the physiological responses of individuals in relation to their personality traits and cognitive load. We preset a detailed explanation of the collected data and descriptive statistics of the results dissected along different measurement dimensions. We then extract correlations among cognitive load measures, physiological signals, and personality traits. Finally, we develop machine learning (ML) models that with up to 82% accuracy (c.f. 50% baseline) predict the cognitive load level experienced by a study participant. The contributions of the work, however, reach beyond this paper as in this work we prepare the datasets for a public release and, based on one of dataset (and our ML models as the baseline), organize a machine learning challenge for cognitive load inference[2].

2. Related Work

Cognitive load represents an important aspect modulating human behavior and a timely and reliable assessment of a person's cognitive load would enable a range of new and improved applications

[1] Link to the two datasets
[2] https://www.ubittention.org/2020/

in areas spanning from game-based learning, over simulator-based driving training, to considerate pervasive human-computer interaction. Yet, the concept remains intangible, thus difficult to grasp and measure. In this section we provide an overview of theoretical postulates behind the concept of cognitive load and recent efforts in measuring cognitive load. In addition, having in mind the nature of this paper, we also provide a brief survey of the existing open datasets in the field of Affective Computing.

2.1. Cognitive load: from theory to measurements

Paas and van Merrienboer define cognitive load as "*a multidimensional construct representing the load that performing a particular task imposes on the learner's cognitive system*" [23]. As such, cognitive load is dependent on the task, the participant, and the interaction between the two. For instance, tasks may be objectively more or less demanding, people can have different cognitive capacities, and certain tasks can be easier for those who are skilled in similar tasks. This multi-dimensionality of cognitive load makes its measurement a rather challenging feat.

Cognitive load measurement methods often rely on data about the subjective perception of the task difficulty, performance data using primary and secondary task techniques, and psycho-physiological data [24]. Measuring subjective data is performed using surveys (e.g. NASA-TLX [25]) solved by a user at the end of a task. However, subjective post-hoc measurements are impractical in real-world applications, as they require explicit querying of users. Cognitive load measurement through a secondary task performance requires a user to attend to a simple secondary task (for instance, react to a slowly changing screen background color), while solving the primary task [26]. These techniques, too, are invasive and in numerous situations, such as while driving, not suitable for in-situ cognitive load inference.

Instead, physiological techniques for cognitive load measurement rely on signals, stemming from heart beat activity [27], breathing [28], heat flux [29], brain activity [30] and eye movement [31,32]. Changes in these signals are a result of our autonomic nervous system's reaction to increased cognitive load. In [29] Haapalainen et al. used elementary cognitive tasks (ECTs), a well-established tool in educational psychology [33], to elicit different levels of cognitive engagement and monitor users' eye movement, heart and brain activity, and skin conductance, while the users are solving the ECTs. The authors demonstrated that two extreme levels of task difficulty ("easy" vs "difficult") can be discriminated by with 80% accuracy using heat flux, ECG features and person-specific data, i.e., personalized models. The method, however, requires that the users are static and strapped with specialized sensors. In a study with developers engaged in real-world programming tasks, Zuger et al. used physiological sensors to infer human interruptibility. The study shows that EEG signals, eye blinks, skin conductance, heart rate and inter-beat interval features correlate with interruptibility, which, in turn negatively correlates with a user's mental load [34].

Recent advancements in sensing technology enable less intrusive forms of vital signs monitoring and get us closer towards unintrusive cognitive load inference [35]. Gjoreski et al. used commercially available Empatica wristbands and acquired signals related to heart rate variability, blood volume pulse, GSR, skin temperature, and acceleration, while exposing users to varying levels of stress [11]. The study demonstrates that off-the-shelf equipment can be used for reliable (up to 92% accuracy achieved in the study) stress detection. While a separate concept, stress may be related to cognitive load, and an earlier study by Setz et al. has already shown that the same GSR sensor can be used to discriminate between the two phenomena [36]. Researchers have also attempted to unobtrusively measure cognitive load in specific environments. For instance, Novak et al. used MS band to infer cognitive load in a simulated driving environment [37]. The authors argue that cheap wearables may provide enough information about physiological signals to enable binary ("engaged in a task" vs "not engaged in a task") classification of the cognitive load, yet are unlikely suitable for inferring the actual level of cognitive load. Schaule et al. use the same wristbands and an N-back task to elicit different levels of cognitive load among office workers [38].

Nevertheless, all the above work treats users as equals, whereas their (physiological) reactions to mental burden might be highly individual. In this paper we rely on the current tendencies of unobtrusive wearables-based cognitive load monitoring, yet, we for the first time introduce personality traits, an important user-level factor impacting cognitive load expression.

2.2. Open datasets for Affective Computing

Open datasets are a staple of reproducible and verifiable science, and may often catalyze significant research activity. Table 1 presents an overview of publicly available datasets in Affective Computing. We particularly focus on datasets that encompass data originating from physiological sensors such as: EEG sensor, electrooculogram (EOG) sensor, electrocardiogram (ECG) sensor, electromyogram (EMG) senso, Blood Volume Pulse Sensor (BVP) sensor, electrodermal activity (EDA) sensor, respiration rate sensor (RESP), eye tracker, Magnetoencephalography (MEG) sensor, skin temperature sensor (TEMP), acceleration sensor (ACC), sensor for beat-to-beat intervals (RR sensor), and Pulse oximetry sensor (SpO2).

The six datasets, Ascertain [39], Amigos [40], DEAP [41], Mahnob [42], Decaf-movies and Decaf-music [43], are emotion recognition datasets where the participants watched affective multimedia in short sessions, e.g., with a duration of 50 to 80 seconds, and rated their experience after each affective session using psychological questionnaires. In all the datasets, the affective multimedia are short movie- or music-video clips designed to induce certain affective states (e.g., fear, surprise, joy, etc.). While the participants were watching the affective multimeda, their physiological response was recorded using a variety of devices. The dataset Emotions differs from the previous datasets as it contains data from a single participant over three weeks, standing in contrast to the studies that examine many participants over a short recording interval. Laughter is another slightly different dataset, which aims at laughter recognition using non-invasive wearable devices.

The three datasets, Driving-workload [44], Driving-stress[45] and Driving-distractions [46] are collected in studies where the main task is driving. In Driving-workload, the participants drove a predefined route including different sections (e.g., crowded vs. free highway) and marked their mental workload afterwards by watching a video recording of the driving session. Similarly, in the Driving-stress dataset, the participants drove different sections and marked the perceived stress level. In addition, this study introduced "a computed stress level" which is calculated based on the situation on the road (e.g., number of cars, pedestrians and signs). The Driving-distractions dataset is a driving-simulator study which analyzes the behaviour of the drivers under different types of stressors (physical, emotional, cognitive and none), and it can be used for development of machine learning models for monitoring driving distractions [47].

The three datasets, Stress-math [11], WESAD [48] and Non-EEG [49] are collected in studies focused on psychological stress. In the Stress-math dataset, the participants were solving simple mathematical questions under time and evaluation pressure. The goal of this study was to induce and recognize psychological stress. In the WESAD dataset, the participants experienced both an emotional and a stress stimulus. More specifically, WESAD contains three sessions for each participant: a baseline session (neutral reading task), an amusement session (watching a set of funny video clips), and a stress session (being exposed to the Trier Social Stress Test [50]). Similarly, Non-EEG is a dataset recorded during three different stress conditions including a physical, a cognitive, and an emotional stressor.

Different to the already available datasets in Affective Computing, this study introduces two new datasets that enable cognitive load monitoring with a wrist-device in combination with personality traits. The Snake dataset is a labeled dataset of cognitive load measurements in which participants were playing a smartphone game. The CogLoad dataset is the first dataset that allows analysis of the cognitive load induced by six different tasks in relation to the physiological responses of individuals and their personality traits. To the best of our knowledge, the only other vaguely related dataset that includes personality traits is the Amigos, which focuses on human emotions.

Table 1. Publicly available dataset from the Related work.

Dataset	Participants	Scenario	Signals
Ascertain	58	Valence, arousal, liking, engagement, familiarity, Big Five	ECG, EDA, EEG, facial activation units
Amigos	40	Valence, arousal, control, familiarity, liking and discrete emotions	EEG, ECG, GSR, face video
DEAP	32	valence, arousal, liking, dominance, and familiarity	ECG, EDA, EEG, EMG, EOG, RESP, TEMP, face video
DECAF-music DECAF-video	30	Valence, arousal, and dominance	ECG, EMG, EOG, MEG, near-infrared face video
Mahnob	30	Valence, arousal, dominance, predictability, and discrete emotions	ECG, EDA EEG, RESP, TEMP, face and body video, eye gazetracker, audio
Emotions	1	Neutral, anger, hate, grief, joy, platonic love, romantic love, reverence	ECG, EDA, EMG, RESP
Laughter	34	Laughter vs. other	ACC, EDA, PPG, TEMP
Driving-work.	10	Perceived cognitive load while driving	GSR, HR, TEMP
Driving-stress	24	Stress levels (low, medium, high)	ECG, EDA, EMG, RESP
Driving-distract.	64	Stress binary + NASA TLX	EDA,HR, RESP, facial expressions, eye tracking
Stress-math	21	Stress levels: (low, medium, high)	ACC, EDA, HR, TEMP, BVP
Non-EEG	20	Four types of stress (physical, emotional, cognitive, none)	ACC, EDA, HR, TEMP, SpO2
WESAD	15	Neutral, amusement, stress	ACC, EDA, TEMP, BVP, EMG, RESP
CogLoad	23	6 different cognitive load tasks. Each with three difficulty levels (easy, medium, hard). Additionally, 2-back and 3-back tasks. NASA-TLX	ACC, EDA, TEMP, RR
Snake	23	Smartphone game with three difficulty levels (easy, medium, hard). NASA-TLX	ACC, EDA, TEMP, RR

2.3. Personality traits, physiological responses and wearables

Research on the relationship between personality traits and physiological responses is not new, and has been done in multiple domains, commonly in research on stress [51,52], aversive stimuli [53,54] and medical issues [55,56]. Most research, however, has not been conducted in order to produce datasets ready for analysis, especially in machine learning. Furthermore, most research is conducted with immovable and expensive instruments for measuring physiological responses. Research with inexpensive wearables for sensing physiological responses that also includes personality assessment and analysis is rare. The likely reason is that the market for such wearables is still new, but also because of the unawareness of the potential of personality traits as input data for ML models. The little research that includes wearables and personality traits so far has mostly focused on emotions [57,58] and stress [59]. We are not aware of research on cognitive load in a similar capacity to ours.

3. Datasets

3.1. CogLoad

In the conducted experiments, the participants were solving cognitive tasks of varying difficulty. The experiments were performed in a quiet, normal-temperature room with one participant at a time. At the beginning of each session, the participants were placed in a comfortable chair in front of a computer monitor and were presented with brief information regarding the experiments. Next, a wrist-device (MS band) was put on their left wrist and the rest of the experimental session was recorded in the same chair without any restrictions regarding the participants' hand gestures. Thus, the experimental setup simulates sedentary work on a computer in an office.

The experimental scenario consists of Part 1 and Part 2. Part 1 was dedicated to assessing the participants' cognitive capacity and the personality type. For assessing the participants' cognitive capacity, the participants were solving two N-back tasks [60], i.e., 2-back and 3-back task, with a three-minute rest after each of them. For assessing the personality type, the participants filled a Hexaco Personality questionnaire, which provides information about the participants': Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience [61].

In Part 2, the participants were presented with six primary tasks. For each task, three variations of a randomly selected primary cognitive-load task were presented to the participant. The variations differed in difficulty (Easy, Medium and Difficult) and thus in the expected cognitive load. After each of the three variations, the participants filled the NASA-TLX questionnaire to assess *subjective* cognitive load posed by the tasks. This questionnaire, the most common means of measuring cognitive load, contains a set of questions that, if administered immediately after the task, allows post-hoc analysis of the cognitive load [25]. The questions identified by the NASA-TLX questionnaire assess mental, physical and time effort, quality of performance, effort, and frustration.

Also, in parallel with the primary tasks, a secondary task was presented to fill-in the participant's free cognitive resources. The secondary task contained a square starting as completely transparent in a random placement of the PC screen, and then increasing in opacity. The participant's goal was to react, i.e., to click on the appearing square as soon as they notice it. The opacity of the square when clicked was intended to be related to the participant's engagement in the primary task, since more engaged users were expected to notice the square later, when it is darker [26]. An assumption is that increased engagement corresponds to higher cognitive load put towards the primary task.

The software, developed by Haapalainen et al. [29] in their study on psycho-physiological measures for assessing cognitive load, was used to display the primary tasks. The software displays the following tasks: Gestalt Completion test - where the participant is asked to identify incomplete drawings; Hidden Pattern test - where the participant has to decide whether a model image is hidden in other comparison images; Finding A's test - where the participant has to find the letter 'a' in presented words; Number Comparison test - Where the participant has to decide whether or not two displayed numbers are the same; Pursuit test - where the participant has to visually track irregularly curved overlapping lines from numbers on the left side of a rectangle to letters on the opposite side; and Scattered X's test - where the participant has to find the letter 'x' on screens containing random letters at random placements. More details about the technical implementation can be found in Novak's thesis [20], while we present the statistical properties of the dataset in *Section 4.1*.

3.2. Snake

A specific version of the game Snake[3] was implemented on Android smartphones. The implemented version allowed varying the difficulty by changing the speed of the game. 23 participants played the game at three difficulty levels: Easy, Medium and Difficult. Each level lasted at least two minutes. Immediately after the completion of each difficulty level, the participants answered a questionnaire to determine perceived difficulty. Difficulty levels were followed with 50% probability in the order from Easy, over Medium, to Difficult, or vice versa. The questionnaire included the NASA Task Load Index (NASA-TLX) questionnaire, plus two general questions about how challenging and fun the game seemed to the user. These questions were answered by the users on 7-point Likert scales across six categories. For assessing the personality type, the participants filled a Hexaco Personality questionnaire.

To assess the participants' physiological response, the MS band wrist-device was used. The data output was: Heart rate, RR intervals, GSR, TEMP, ACC data. The HR, the GSR and the TEMP were sampled at 1 Hz, the ACC was sampled at 8 Hz, and the RR intervals were recorded upon detection (e.g., for 60 beats-per-minute, the frequency would be 1Hz). Additionally, the screen-tapping speed was recorded. The data was transmitted via Bluetooth from the wrist-device to the smartphone and then to a server. More details about the technical implementation can be found in Knez's thesis [19], while we describe the statistical properties of the dataset are described in *Section 4.1*.

4. Psychological and Behavior Analysis

Multi- and interdisciplinary efforts in computer science towards combining heterogeneous data in understanding and predicting targets related to complex cognitive phenomena with the help of computational methods, especially machine learning, are bearing fruit in discovering that physiological and psychological data interact in beneficial ways. Performing descriptive and similar statistics on psychological data, as is the norm in psychological, behavioral and cognitive sciences, therefore has a place in primarily computer science fields as well.

This section presents various statistical analyses of demographic, psychological and cognitive load data from the two datasets. It uses them to discuss on the reasons for various correlations and other factors relating to performance and cognitive load results. This is mostly to create a baseline demonstration for how demographic and psychological data can be exploited. *Section 5* discusses more advanced analyses and interpretations. A detailed interpretation of the presented statistics is provided in *Section 6.1*

4.1. Personality – descriptive analysis of the datasets

The CogLoad dataset includes 23 randomly selected participants, sampled in Slovenia. Participants' mean age was 29.51 (standard deviation being 10.10), and their highest attained education levels were as follows: a high school diploma in 7 cases (30.43%), a bachelor's degree in 6 cases (26.09%), a master's degree (26.09%) in 6 cases, and a doctoral degree in 4 cases (17.39%). Right was the dominant hand of 2 participants, while 1 participant was left-handed. All participants had the MS Band device strapped to their left hand. The Snake dataset includes 23 (16 men and 7 women)

[3] https://en.wikipedia.org/wiki/Snake_(video_game_genre)

randomly selected participants, sampled in Slovenia. Participants' mean age was 24.91 (standard deviation being 12.05). The HEXACO personality questionnaire was administered with each of the participants in both datasets.

The personality analysis (descriptive statistics and correlations) we present here comes from the HEXACO questionnaire, which is based on six factor-level (higher level) scales or dimensions, each separated into lower facet-level scales. The six factor-level scales with multiple facet-level scales include:

1. **Honesty-Humility** measures: *Sincerity*, *Fairness*, *Greed Avoidance*, *Modesty*.

   People that rank high on Honesty-Humility do not pursue personal gain to the others' detriment, they follow the rules, they do not seek large material wealth, and do not judge people by their social status. On the opposite side of the spectrum, people that rank low on Honesty-Humility are prone to manipulating people, breaking rules, seeking material wealth over other goals, and feeling more important than others.
2. **Emotionality** measures: *Fearfulness*, *Anxiety*, *Dependence*, *Sentimentality*.

   People that rank high on Emotionality are extremely fearful of physical dangers, they are very prone to feel anxious when under stress, they constantly seek external support, and are very empathetic. On the opposite side of the spectrum, people that rank low on Emotionality easily overcome fear of physical dangers, they do not worry a lot even when under stressful duress, they quickly find internal support for their matters, and detach from others emotionally.
3. **Extraversion** measures: *Social Self-Esteem*, *Social Boldness*, *Sociability*, *Liveliness*.

   People that rank high on Extraversion have high self-esteem, they are confident, they are often leadership material, they feel comfortable at social events, and are enthusiastic. On the opposite side of the spectrum, people that rank low on Extraversion are self-conscious, they cannot manage being the center of attention, they do not enjoy social gatherings, and are generally less optimistic.
4. **Agreeableness** measures: *Forgivingness*, *Gentleness*, *Flexibility*, *Patience*.

   People that rank high on Agreeableness quickly forgive people, they do not judge people, they have no problems cooperating with other people, and they manage their anger well. On the opposite side of the spectrum, people that rank low on Agreeableness, often hold grudges towards others for long periods of time, they are fast to criticize, they are not easily convinced they are wrong, and they react with anger in many situations.
5. **Conscientiousness** measures: *Organization*, *Diligence*, *Perfectionism*, *Prudence*.

   People that rank high on Conscientiousness are great at organizing their time and space, they can plan well towards their short-, medium- and long-term goals, they are precise and can be perfectionists, and always take time to think on their courses of action. On the opposite side of the spectrum, people that rank low on Conscientiousness do not bother with having or respecting schedules, they prefer leisure to challenge, they are quickly satisfied in whatever they do, and act spontaneously and without thought.
6. **Openness to Experience** measures: *Aesthetic*, *Inquisitiveness*, *Creativity*, *Unconventionality*.

   People that rank high on Openness to Experience are fascinated by aesthetics, be it in art or nature, they are extremely eager to learn, they use imagination in every aspect of their lives, and they are attracted to that which is out of the norm. On the opposite side of the spectrum, people that rank low on Openness to Experience are not interested in aesthetics, they do not pursue knowledge, they lack creativity, and are fine with conforming.

Table 2. Personality scores from the HEXACO questionnaire.

	CogLoad dataset		Snake dataset		L&A (2016)	
	M	*SD*	*M*	*SD*	*M*	*SD*
Honesty-Humility	3.29	0.60	3.32	0.67	3.30	0.74
Emotionality	2.91	0.75	3.19	0.53	3.12	0.63
Extraversion	3.12	0.68	2.97	0.73	3.22	0.64
Agreeableness	3.11	0.59	3.24	0.57	2.78	0.63
Conscientiousness	3.48	0.57	3.10	0.66	3.52	0.55
Openness	3.43	0.79	3.10	0.61	3.69	0.57

For the CogLoad dataset, factor-level and facet-level scales were calculated from the questionnaire answers. For the Snake dataset, only factor-level scales were calculated. Table 2 shows the mean (*M*) and the standard deviation (*SD*) of our sample from the CogLoad and Snake datasets. No division into further groups (sex, age, education, handedness) was performed due to the low *N*. The table also shows *M* and *SD* of 100,318 self-reports from [62] for comparison purposes ('L&A (2016)' label in the Table).

4.2. Personality, TLX, and objective cognitive load analysis

The data on psychological traits, TLX scores and objective cognitive load was used for this analysis. Due to the high variation in 95% Confidence Interval scores, all correlations are presented in the tables. We are aware that commonly, only correlations with a minimum inclusion threshold of 0.3 in absolute value are presented, as such a correlation denotes a medium or higher (strong) correlation strength [63], while below 0.3 correlation is considered as weak. Spearman correlation was used for the presented scores for higher robustness.

Table 3 presents the correlations of medium and above strength between personality traits and selected dimensions of the TLX scores for the CogLoad dataset with 95% Confidence Interval in parentheses. The label 'TLX_physical_demand' represents a score on the questions "How much physical activity was required?" and "Was the task easy or demanding, slack or strenuous?". Emotionality is a factor-level trait, while Dependence, Fearfulness and Anxiety are Emotionality's facet-level traits.

Table 3. Correlations between personality traits and the TLX-scores for the CogLoad dataset with 95% Confidence Interval in parentheses.

	emotionality	dependence	fearfulness	anxiety
TLX_physical_demand	+0.523 (0.14-0.77)	+0.470 (0.07-0.74)	+0.386 (-0.03-0.69)	+0.380 (-0.04-0.68)

Table 4 presents the correlations between the TLX scores and objective cognitive load measures for the CogLoad dataset with 95% Confidence Interval in parentheses. The label 'time_on_task' represents the time a participant spent on a task; 'num_correct' represents the number of correct answers; 'level' represents the difficulty level of the task; 'TLX_mean' represents the average of all TLX scores; 'TLX_effort' represents a score on the question "How hard did you have to work (mentally and physically) to accomplish your level of performance?"; 'TLX_temporal_demand' on "How much time pressure did you feel due to the pace at which the tasks or task elements occurred?" and "Was the pace slow or rapid?"; 'TLX_mental_demand' on "How much mental and perceptual activity was required?" and "Was the task easy or demanding, simple or complex?"; 'TLX_frustration' on "How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?"; and 'TLX_performance' on "How successful were you in performing the task? How satisfied were you with your performance?".

Table 5 presents the correlations between personality traits and the objective cognitive load measures for the Snake dataset with 95% Confidence Interval in parentheses. The label 'Points' represents the number of points the participant got while playing the snake game.

Table 4. Correlations between the TLX-scores and objective cognitive load measures for the CogLoad dataset with 95% Confidence Interval in parentheses.

	time_on_task	**num_correct**	**temperature**	**level**
TLX_mean	+0.503 (0.11-0.76)	-0.286 (-0.62-0.14)	+0.345 (-0.08-0.66)	+0.274 (-0.16-0.62)
TLX_effort	+0.490 (0.10-0.75)	-0.289 (-0.63-0.14)	+0.263 (-0.17-0.61)	+0.282 (-0.15-0.62)
TLX_temporal_demand	+0.484 (0.09-0.75)	-0.176 (-0.55-0.25)	+0.275 (-0.16-0.62)	+0.161 (-0.27-0.54)
TLX_mental_demand	+0.419 (0.01-0.71)	-0.266 (-0.61-0.16)	+0.335 (-0.09-0.66)	+0.258 (-0.17-0.61)
TLX_frustration	+0.365 (-0.06-0.68)	-0.185 (-0.55-0.25)	+0.350 (-0.07-0.67)	+0.125 (-0.3-0.51)
TLX_performance	+0.346 (-0.08-0.66)	-0.325 (-0.65-0.10)	+0.117 (-0.31-0.50)	+0.135 (-0.29-0.52)
TLX_physical_demand	+0.127 (-0.3-0.51)	-0.023 (-0.43-0.39)	+0.416 (0.00-0.71)	+0.033 (-0.38-0.44)

Table 5. Correlations between personality traits and the objective cognitive load measure for the Snake dataset with 95% Confidence Interval in parentheses.

	Heart rate	**Temperature**	**Points**
Openness	-0.017 (-0.43-0.40)	-0.165 (-0.54-0.27)	+0.347 (-0.08-0.66)
Conscientiousness	+0.093 (-0.33-0.49)	-0.252 (-0.60-0.18)	+0.193 (-0.24-0.56)
Extraversion	-0.172 (-0.55-0.26)	-0.185 (-0.55-0.25)	+0.059 (-0.36-0.46)
Honesty-Humility	-0.201 (-0.57-0.23)	-0.026 (-0.43-0.39)	+0.136 (-0.29-0.52)
Emotionality	+0.336 (-0.09-0.66)	-0.056 (-0.46-0.36)	+0.016 (-0.40-0.43)
Agreeableness	-0.115 (-0.50-0.31)	+0.316 (-0.11-0.64)	+0.300 (-0.13-0.63)

Table 6 presents the correlations between the TLX-scores and objective cognitive load measures for the Snake dataset with 95% Confidence Interval in parentheses. Label 'subjective diff' represents the subjective score of how difficult the game was, 'level' represents the game's difficulty level, 'click per second' represents the number of clicks the participant made during the measuring time, 'gsr' represents the galvanic skin response, 'hr' represents the heart rate, and 'TLX_effort' represents a score on the question "How hard did you have to work (mentally and physically) to accomplish your level of performance?".

Table 6. Correlations between the TLX-scores and the objective cognitive load measure for the Snake dataset with 95% Confidence Interval in parentheses.

	subjective diff	**level**	**click per second**	**temperature**	**gsr**	**hr**
subjective diff	1	+0.742 (0.47-0.88)	+0.520 (0.14-0.77)	+0.035 (-0.38-0.44)	-0.009 (-0.42-0.4)	+0.062 (-0.36-0.46)
level	+0.742 (0.47-0.88)	1	+0.595 (0.24-0.81)	-0.032 (-0.44-0.39)	-0.045 (-0.45-0.37)	+0.072 (-0.35-0.47)
TLX_effort	+0.818 (0.61-0.92)	+0.648 (0.32-0.84)	+0.375 (-0.04-0.68)	+0.077 (-0.35-0.47)	+0.035 (-0.38-0.44)	-0.008 (-0.42-0.41)
TLX_mental demand	+0.750 (0.49-0.89)	+0.513 (0.13-0.76)	+0.319 (-0.11-0.65)	+0.075 (-0.35-0.47)	+0.062 (-0.36-0.46)	-0.006 (-0.42-0.41)
TLX_temporal demand	+0.669 (0.35-0.85)	+0.671 (0.36-0.85)	+0.459 (0.06-0.73)	+0.053 (-0.37-0.46)	+0.038 (-0.38-0.44)	+0.105 (-0.32-0.5)
TLX_physical demand	+0.266 (-0.16-0.61)	+0.181 (-0.25-0.55)	-0.042 (-0.45-0.38)	-0.539 (-0.78-(-0.16))	-0.456 (-0.73–0.05)	-0.304 (-0.64-0.12)
TLX_performance	-0.383 (-0.69-0.04)	-0.353 (-0.67-0.07)	-0.261 (-0.61-0.17)	-0.132 (-0.52-0.30)	-0.018 (-0.43-0.4)	+0.014 (-0.4-0.42)
TLX_frustration	+0.413 (0.00-0.70)	+0.385 (-0.03-0.69)	+0.144 (-0.29-0.52)	-0.474 (-0.74-(-0.08))	-0.402 (-0.7-0.01)	-0.149 (-0.53-0.28)

5. Machine Learning Analysis

In this section we present a suite of machine learning modeling approaches that connect the data sensed by the Microsoft Band wristband with the outcome, i.e. the experienced level of cognitive load. Having in mind the susceptibility of subjective metrics of cognitive load to interpretation (potentially modulated by a participant's personality), here we focus on the objective/designed difficult of a task, and binary Easy/Hard classification as explained in *Section 5.3*.

5.1. Preprocessing, segmentation and feature extraction

We initially re-sampled all the data to a sampling frequency of 1 Hz. Next, the last 30 seconds of each task were used to extract features. Thus, one segment represents one task. For each segment, statistical features were extracted from each input signal i.e., Heart rate, RR intervals, GSR and TEMP, and their first differentials. The statistical features include: mean, standard deviation, skewness, kurtosis, mean of the first derivative, mean of the second derivative, 25th and 75th percentile, inter-quartile range, difference between the minimum and the maximum values and coefficient of variation.

Additional features were extracted from the GSR signal using Skin Conductance Response (SCR) analysis. This type of features/analysis is proven to be useful for detecting stressful conditions in driving scenarios [45] and in real-life situations [11]. The GSR signal is first preprocessed using a sliding mean filter, and then fast-acting (GSR responses) component and slow-acting (tonic) component were extracted. The fast-acting component was used to calculate the number of responses in the signal, the responses per minute in the signal, and the sum of the responses. The slow-acting component was used to calculate the mean value of the first differentials of the tonic component, and the difference between the tonic component and the overall signal.

The activation of the sympathetic nervous system triggered by cognitive load leads to more equidistant heart beats. On the other hand, the rest periods between the tasks reverse this process, and the heart beats become more irregular, as "A healthy heart is not a metronome" [64]. Heart Rate Variability Analysis (HRV) is commonly used to quantify the dynamics of the RR intervals. The RR signal was filtered by removing the outliers, i.e., the RR intervals that are outside of the interval [0.7*median, 1.3*median], where the median is segment-specific. Next, the following HRV features were calculated: the mean heart rate, the standard deviation of the RR intervals, the standard deviation of the differences between adjacent RR intervals, the square root of the mean of the squares of the successive differences between adjacent RR intervals, the percentage of the differences between adjacent RR intervals that are greater than 20 ms, the percentage of the differences between adjacent RR intervals that are greater than 50 ms, and Poincare plot indices (SD1 and SD2) [65].

5.2. Normalization, feature selection and model learning

To analyze the inter-participant and inter-session influence, experiments were performed without normalization, with session-specific min-max normalization and with session-specific standardization. When min-max normalization is used, each feature is scaled between 0 and 1 by subtracting the minimal value and then by dividing this difference with the difference between the minimal and the maximal values. When standardization is used, each feature is mean centered by subtracting the mean value and then dividing with the standard deviation.

Additionally, experiments were performed with and without feature selection. In general, all feature selection methods can be divided into wrapper methods, ranking methods (also known as filter methods) and a combination of the two. The wrapper methods (e.g., based on ROC metrics [66]) produce better results compared to the ranking methods (e.g., information entropy [67]), but they induce a heavy computation burden. In this study, ranking method based on mutual information [68] was used, because it is very efficient to compute. Mutual information is a measure that estimates the dependency between two random variables. The features were ranked using mutual information values between the features and the class values estimated on the training data, and only the top-ranked 50 features were used to build models.

Experiments were performed with the following ML algorithms: Decision Tree [69], RF [70], Naïve Bayes [71], KNN [72], Logistic Regression [73], Bagging - using Decision Trees [74], Gradient Boosting (AdaBoost), Extreme Gradient Boosting (XGB) and Multi-layer perceptron (MLP) [75]. The specific architecture used for the MLP is available online[4]. It contains two hidden layers, one of size 512 and one of size 32 units, and one output unit that uses sigmoid activation function.

These ML algorithms learn one model for each training dataset. The ML approach capable of learning models for several ML datasets (ML tasks) in parallel while using a shared representation is Multi-task learning (MTL) [76]. The idea is to use what was learned from one dataset to help learn other tasks better. More specifically, in single-task neural networks, backpropagation algorithm is used to minimize a single loss function and single neuron provides the final output. MTL, on the other hand, involves the minimization of a joint loss function (e.g., weighted sum of the binary cross-entropies of all tasks) and learning shared representations over all tasks (see Figure 1). The specific MTL architecture was similar to the MLP architecture[5]. It contains two shared-hidden layers of size 512 units, one task specific layers of size 32 units and two task-specific sigmoid units that output the final predictions.

Both for the MTL and MLP architectures, ReLU activation units [77] were used in the hidden layers, which speeds up the training process compared to other activation layers (e.g., tanh). To avoid overfitting, L2 regularization and dropout was used. The training of the networks was fully supervised, by back propagating the gradients through all the layers. The parameters were optimized by minimizing the binary cross-entropy loss function using the Adam optimizer. The models were trained with a learning rate of 10^{-4} and a decay of 10^{-4}. The batch size was set to 32 and the number of training epochs was set to 50.

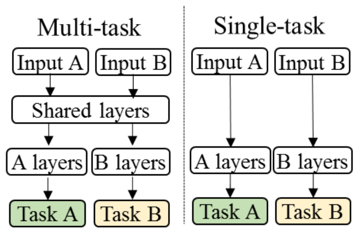

Figure 1. Multi task learning vs. Single task learning.

5.3. Experimental setup

Leave-one-session-out evaluation techniques was used in all ML experiments. This means that the data of one session was used as a test data, and the rest of the data was used for training and tuning the ML models. In the CogLoad dataset, there is only one session per participant, thus the models are participant-independent. In the Snake dataset, there are more than one sessions for some participants, thus the models are participant-dependent.

For each ML algorithm, parameter tuning was performed using the following procedure: parameter settings were randomly sampled from distributions predefined by an expert. Next, models were trained with the specific parameters and then evaluated using internal k-fold cross-validation on the training data. The best performing model from the internal k-fold cross-validation was used to classify the test data. This tuning procedure was performed as many times as there were sessions in the specific experimental dataset. Additionally, the evaluation was repeated five times to account for the randomness present in the learning (e.g., Random Forest) and the tuning (e.g., the random parameter sampling) of the ML models.

For the CogLoad dataset, the ML task was the classification of rest vs. task segments. For the Snake dataset, the ML task was the classification of easy vs. hard segments. The rest periods were not recorded in the Snake dataset, thus rest vs. task classification is not possible. Additionally, for the Snake dataset, the segments with medium difficulty were not used in the ML analysis following the

---

4 https://repo.ijs.si/martingjoreski/cognitive-load/-/blob/master/MLP.png
5 https://repo.ijs.si/martingjoreski/cognitive-load/-/blob/master/MTL.png

studies by Rissler et al. [78] and Maier et al. [79], in which only the top 20% and the lowest 20% of the data points were considered for the classification task. The data points that fall in between were discarded. Table 7 presents the size of the experimental datasets after the labelling. Each instance represents a 30-second segment labelled with a "High" or "Low" difficulty.

The averaged results for a binary classification problem are presented in Table 8. All models were dataset-specific, except for the MTL model, which is a joint model for the two datasets. The last three columns present the accuracy of the ML models built using selected features in combination with raw features (without any normalization), normalized features (min-max normalization) and standardized features. The three columns before that present the accuracy of each ML model built using all features in combination with raw features, normalized features, or standardized features.

Table 7. Number of instances in the ML experiments for the two datasets.

	Low	**High**	**Overall**
Snake	34	35	69
CogLoad	412	413	825

Table 8. Machine learning evaluation results with binary classification accuracy.

Dataset	**Model**	**All features**			**Selected features**		
		Raw	**Norm.**	**Stand.**	**Raw**	**Norm.**	**Stand.**
CogLoad	Majority	50.0	50.0	50.0	50.0	50.0	50.0
	Random Forest	62.4	64.9	66.8	62.9	64.5	67.9
	AdaBoost	60.4	64.3	65.6	61.8	61.7	67.3
	KNN	51.7	59.3	63.6	58.2	59.8	64.0
	Naive Bayes	49.0	63.3	58.5	51.3	60.8	57.9
	Decision Tree	58.6	61.8	63.9	59.6	60.4	62.0
	Log. Reg.	60.3	62.5	64.0	61.1	63.7	65.7
	Bagging	63.9	65.0	67.4	64.7	65.2	**68.2**
	XGB	61.6	63.1	65.5	62.2	61.9	66.4
	MTL	63.3	64.1	63.4	63.6	64.0	65.2
	MLP	63.7	62.2	62.8	63.9	64.3	63.1
Snake	Majority	51.0	51.0	51.0	51.0	51.0	51.0
	Random Forest	60.0	67.7	68.3	61.4	66.9	70.0
	AdaBoost	57.7	66.0	66.6	54.6	69.4	70.3
	KNN	52.3	70.6	68.0	57.1	70.6	67.7
	Naive Bayes	51.4	67.1	68.6	52.9	66.0	68.6
	Decision Tree	55.4	66.0	64.6	56.0	68.3	71.4
	Log. Reg.	56.0	67.7	68.6	56.9	66.3	70.6
	Bagging	60.0	68.0	67.7	60.9	73.1	69.4
	XGB	59.1	79.1	82.0	58.0	78.3	**82.3**
	MTL	69.5	58.6	65.7	71.4	67.1	70.0
	MLP	60.0	61.0	63.8	62.4	62.9	65.7

6. Discussion

6.1. Results discussion

The discussion examines the relationships between personality, cognitive load measures, and physiological data. To the best of our knowledge, this is the first research that tries to examine such results and interpret them. The examination focuses on correlations with at least medium correlation strength (+/-0.3 as the threshold).

Table 3 shows significant correlations between personality traits and physical demand as measured by TLX for the CogLoad dataset. Emotionality and its three facet-level traits, Dependence, Fearfulness and Anxiety, all significantly positively correlate with subjective physical demand. Since Emotionality describes a response to stressful and demanding situations as well as physical danger, the positive correlation is sensible, meaning that people that rank high in Emotionality also find tasks physically more demanding, and vice versa.

Table 4 shows significant correlations between the TLX scores and objective cognitive load measures for the CogLoad dataset. The correlations show that people that spend more time on tasks find them more demanding, put in more effort, get more frustrated, but also feel they performed better the more time they spent on them. The negative correlation between the correct answers and perceived performance, however, is unusual. It reports that the more people felt they did well, the worse they actually performed. Whether this is due to chance or a measurement problem is unclear, but should be noted as something to be aware of. Deeper psychological profile construction and comparison could yield possible answers for this correlation. It could be that people that score low on Humility overwhelmingly report higher performance scores, but are very susceptible to cognitive load. That many TLX scores significantly correlate with the task difficulty level is also expected – this mostly confirms that the difficulty levels are reasonably set as they are.

Table 5 shows significant correlations between personality traits and objective cognitive load measures for the Snake dataset. People higher in Emotionality have higher heart rates during solving tasks (and vice versa); people higher in Agreeableness have higher temperature during solving tasks, which is the opposite of expected. The ability of agreeable people is to control temper, and as our temperature rises if we cannot control temper (which can be a response to stress), the correlation should be negative, not positive. This is another result that should be noted for future investigation. Otherwise, more open people and more agreeable people score better. More open people are more skilled in solving complex tasks, which makes this results sensible. More agreeable people are more in control of their frustration, which could result in more points as they have easier time staying focused on the task.

Table 6 shows significant correlations between the TLX scores and other cognitive load measures for the Snake dataset. The results are mostly sensible: subjective difficulty correlates positively with TLX scores, except perceived performance, which is again sensible, as higher difficulty means worse perceived performance. Similar interpretation can be made for the objective level of difficulty as well as clicks per second (as more clicks are usually needed in tasks that have higher difficulty) and their significant correlations. More puzzling are the remaining correlations with temperature, galvanic skin response and heart rate. The more demanding people perceive tasks to be and the more they get frustrated, the lower temperature they have. As discussed before, both should be positively correlated with TLX scores. Same goes for heart rate. The only explanation, if the correlations are causations, can be found in more thorough psychological profiles. It may be that our participants' profiles are such that demanding situations make them focus, thus lowering heart rate, galvanic skin response and temperature. Interpreting correlations is always a difficult, sometimes questionable practice. Here, another presupposition is made before interpretation – that psychological and cognitive data are grounded in more physiological and neural phenomena. Regardless, the discussion shows that there are relationships between such heterogeneous data.

Table 8 shows the ML results. It can be seen that in general, the models perform better on the Snake dataset compared to the CogLoad datases. This is because the CogLoad models are person independent and the Snake models are person dependent. The highest accuracy of 82.3% on the Snake dataset is achieved by the XGB algorithm in combination with feature selection and feature standardization. The highest accuracy of 68.2% on the CogLoad dataset is achieved by the Bagging algorithm in combination with feature selection and feature standardization. Another observation is that the ensemble models (e.g., RF, Bagging and XGB) perform better compared to the single-model algorithms. This is because the ensemble models are more robust to noise. Finally, it should be noted that our ML modeling Was successful only with the two-class version of the cognitive load inference problem (e.g. discerning between Low and High load). A more fine-grained Low/Medium/High load inference proved to be prohibitively difficult for our algorithms, thus was not discussed in this paper.

Regarding the proposed MTL approach, it is interesting to note that for the dataset that contains more instances (the CogLoad dataset), both the MTL and the MLP perform similar. However, for the smaller dataset, the MTL approach consistently outperformed the MLP approach. This may indicate that combining similar datasets using MTL is useful when the target dataset is small.

6.2. Related-work Discussion

A direct comparison with results from the related work is not possible because of the many differences in the experimental setup. The differences include: different datasets, different sensors, different preproceessing steps, different ML methods, different classification tasks, different evaluation procedure, etc. To provide some intuition, Table 9 presents the F1-scores achieved in the studies on emotion recognition. These studies analyze participants' physiological changes induced by a subtle stimuli (e.g., a video), which is similar to our study. All datasets are balanced, i.e., the majority class is close to 50% and all studies perform binary classification tasks (e.g., low vs. high arousal), which means that F1-scores and Accuracy measures provide similar numbers. It can be seen that our results are comparable to the related work. Moreover, it can be seen that building accurate ML models to recognize changes induced by subtle stimuli is challenging. The challenge is even bigger when only a single wrist-device is used. This was also confirmed by Maier et al. [79] in their study for detecting optimal user experience using a wrist-device in participants that played the game Tetris. Their state-of-the-art deep neural network achieved an accuracy of 67.5% in a binary classification problem (high vs. low flow). Haapalainen et al. [29] achieved an average accuracy of 80% for binary classification problem ("easy" vs "difficult" tasks) using personalized ML models and a combination of heat flux and ECG features, derived from specialized sensor equipment. The person-dependent models in this study, achieved similar results using only a wrist-device. The study revealed that the task type and the chosen cognitive load metric on the models' accuracy. However, classifying task difficulty with an accuracy over 80%, on an ML task where the majority class is close to 50%, using person-independent models and unobtrusive sensors is still an open research question. This was also confirmed in our previous study related to the CogLoad dataset, where both task difficulty and TLX scores were used as ground-truth for ML models [80].

Table 9. F1-score achieved in the ML experiments from the related work.

	Ascertain	Amigos	DEAP	Decaf Movie	Decaf Music
Low vs. High Valence	68	57	61	59	59
Low vs. High Arousal	59	57	62	54	55

6.3. Real-life Applications and Limitations

There are many use-cases for the presented datasets and models to enable improvement of meaningful life outcomes. Lohani et al. [81] presented an overview of the psychophysiological measures that can be utilized to assess cognitive states while driving. The psychophysiological measures included: EEG, optical imaging, heart rate and HRV, blood pressure, GSR, ECG, thermal imaging, and pupillometry. Another use-case includes measuring workload of pilots. For example, Mohanavelu et al. [82] analyzed HRV features for measuring cognitive workload of 20 fighter aircraft pilots in a flight simulator environment. The statistical analysis in their study revealed a strong significant difference between workload with respect to HRV parameters. Johannessen et al. [83] analyzed cognitive load in five physician team leaders during trauma resuscitation. Eye-tracking, GSR, and heart rate measures were captured during trauma resuscitations in a real-world setting. Fritz et al. [84] used psycho-physiological measures to assess task difficulty in software development. They conducted a study with 15 professional programmers to see how well an eye-tracker, an GSR, and an EEG sensor could be used to predict whether developers would find a task to be difficult. Jimenez-Molina et al. [85] explored PPG, EEG, temperature and pupil dilation sensors to assess mental workload of 61 participants during web browsing. They evaluated Multinomial Logistic Regression, SVM, and MLP models using 70%-30% train-test split. The best signal modality was EEG with an accuracy of 70%, while the rest of the modalities achieved an accuracy around 35%.

The size of the datasets used in our study is comparable to the related studies on cognitive load [82–84,86–88]. However, the findings should be confirmed in a larger study with more participants, in order to draw general conclusions.

Finally, the secondary task used in the CogLoad dataset may be problematic for participants with vision problems. Any individual differences here could have skewed results. In future similar studies, vision should be taken into account.

7. Conclusion and Future Work

This study presented two datasets of multimodal data sensed with a commodity wearable device, while the participants were exposed to a varying cognitive load. To the best of our knowledge these are the first datasets that include such rich sensor data augmented with the information on the personality traits of the participants. The experimental setup in which the datasets were collected included a variety of cognitive tasks performed on a smartphone and on a PC. We also presented the analysis of the psychological data in relation to the subjective cognitive load (NASA-TLX) and the objective cognitive load measures, revealing potentially significant relationships. For example, we found that people who rank high in Emotionality find tasks physically more demanding and have higher heart rates during task solving (and vice versa). Also, there were evidence that people that score low on Humility may report higher performance scores, but are very susceptible to cognitive load. Furthermore, we present baseline ML models for recognizing task difficulty. The person-independent models on the CogLoad dataset achieved an accuracy of 68.2%, while the person-dependent models on the Snake dataset achieved an accuracy of 82.3%. These results are in line with the related work that uses more sophisticated lab-based measurement equipment. The proposed multi task learning (MTL) neural network outperformed the single task neural network (a multi layer perceptron - MLP) by simultaneously learning from the two datasets. The datasets will be made publicly available to advance the field of cognitive load inference using commercially available devices.

Our next step will be to build ML models that combine both the psychological and physiological data for inferring cognitive load [89]. Personality grouping shows differences between people on a more fundamental level, and these differences can express physiologically. Grouping can be made either through unsupervised learning, i.e., clustering, or expert techniques (e.g., making groups on dominant dimensions). Finding 'noisy' participants is important as well. 1/6 of participants give false answers to psychological questionnaires [90]. For example, in our data, these individuals could be filtered out through the Honesty-Humility trait score. Making separate models for different groups is therefore viable as well. This should improve our current results as well as strengthen our vision for more interdisciplinary research on cognitive phenomena.

Author Contributions: conceptualization, V.P., M.L., M.G, H.G.; methodology, M.Gj., T.K. and V.P.; software, M.Gj., T.K. and T.Kn.; validation, M.Gj., T.K. and T.Kn.; formal analysis, M.Gj., T.K. and T.Kn.; investigation, M.Gj., T.K. and Kn.; resources, all authors; data curation, M.Gj., T.K. and T.Kn.; writing–original draft preparation, all authors; writing–review and editing, all authors; visualization, M.Gj. and T.K.; supervision, V.P., M.L., M.G, H.G; project administration, V.P., M.L., M.G., H.G.; funding acquisition, V.P., M.L., M.G

Acknowledgments: The authors thank Martin Frlin for his assistance in the data collection and the anonymous reviewers for the insightful comments. This work was supported by the Slovenian Research Agency (ARRS) under Grant U2-AG-16/0672, 0287 and Grant N2-0136.

References

1. Peter, C.; Beale, R. *Affect and emotion in human-computer interaction: From theory to applications*; Vol. 4868, Springer Science & Business Media, 2008.
2. Kolenik, T.; Gjoreski, M.; Gams, M. Designing an intelligent cognitive assistant for behavior change in mental health. Slovenian Conference on Artificial Intelligence: proceedings of the 22nd International Multiconference Information Society - IS 2019. Jožef Stefan Institute, 2019, pp. 69–72.

3. Borst, J.P.; Taatgen, N.A.; van Rijn, H. What Makes Interruptions Disruptive?: A Process-Model Account of the Effects of the Problem State Bottleneck on Task Interruption and Resumption. CHI '15. ACM, 2015.
4. Czerwinski, M.; Horvitz, E.; Wilhite, S. A Diary Study of Task Switching and Interruptions. CHI '04. ACM, 2004.
5. Iqbal, S.T.; Horvitz, E. Notifications and Awareness: A Field Study of Alert Usage and Preferences. CSCW '10. ACM, 2010.
6. Mark, G.; Voida, S.; Cardello, A. "A Pace Not Dictated by Electrons": An Empirical Study of Work Without Email. CHI '12. ACM, 2012.
7. Stothart, C.; Mitchum, A.; Yehnert, C. The attentional cost of receiving a cell phone notification. *Journal of experimental psychology: human perception and performance* **2015**.
8. Goyal, N.; Fussell, S.R. Intelligent Interruption Management Using Electro Dermal Activity Based Physiological Sensor for Collaborative Sensemaking. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2017**, *1*, 52:1–52:21. doi:10.1145/3130917.
9. Mark, G.; Gudith, D.; Klocke, U. The cost of interrupted work: more speed and stress. Proceedings of the SIGCHI conference on Human Factors in Computing Systems, 2008, pp. 107–110.
10. Kushlev, K.; Proulx, J.; Dunn, E.W. " Silence Your Phones" Smartphone Notifications Increase Inattention and Hyperactivity Symptoms. Proceedings of the 2016 CHI conference on human factors in computing systems, 2016, pp. 1011–1020.
11. Gjoreski, M.; Luštrek, M.; Gams, M.; Gjoreski, H. Monitoring stress with a wrist device using context. *Journal of biomedical informatics* **2017**, *73*, 159–170.
12. Fredericks, T.K.; Choi, S.D.; Hart, J.; Butt, S.E.; Mital, A. An investigation of myocardial aerobic capacity as a measure of both physical and cognitive workloads. *International Journal of Industrial Ergonomics* **2005**, *35*, 1097–1107.
13. Shakouri, M.; Ikuma, L.H.; Aghazadeh, F.; Nahmens, I. Analysis of the sensitivity of heart rate variability and subjective workload measures in a driving simulator: the case of highway work zones. *International journal of industrial ergonomics* **2018**, *66*, 136–145.
14. Grassmann, M.; Vlemincx, E.; von Leupoldt, A.; Mittelstädt, J.M.; Van den Bergh, O. Respiratory changes in response to cognitive load: A systematic review. *Neural plasticity* **2016**, *2016*.
15. Haak, M.; Bos, S.; Panic, S.; Rothkrantz, L. Detecting stress using eye blinks and brain activity from EEG signals. *Proceeding of the 1st driver car interaction and interface (DCII 2008)* **2009**, pp. 35–60.
16. ElKomy, M.; Abdelrahman, Y.; Funk, M.; Dingler, T.; Schmidt, A.; Abdennadher, S. ABBAS: an adaptive bio-sensors based assistive system. Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems, 2017, pp. 2543–2550.
17. Shi, Y.; Ruiz, N.; Taib, R.; Choi, E.; Chen, F. Galvanic skin response (GSR) as an index of cognitive load. CHI'07 extended abstracts on Human factors in computing systems, 2007, pp. 2651–2656.
18. Wang, Z.; Yan, J.; Aghajan, H. A framework of personal assistant for computer users by analyzing video stream. Proceedings of the 4th workshop on eye gaze in intelligent human machine interaction, 2012, pp. 1–3.
19. Knez, T. Analyzing effect of computer game difficulty on biological signals: Bachelor's Thesis: First Cycle University Study Programme Computer and Information Science, 2019.
20. Novak, M.F. Developing Software Tools for in Situ Cognitive Load Estimation: Bachelor's Thesis: First Cycle University Study Programme Computer and Information Science, 2019.
21. Paunonen, S. Big Five factors of personality and replicated predictions of behavior. *Journal of Personality and Social Psychology* **2003**, *84*, 411–424.
22. Yan, L.; Wang, Y.; Ding, C.; Liu, M.; Yan, F.; Guo, K. Correlation Among Behavior, Personality, and Electroencephalography Revealed by a Simulated Driving Experiment. *Frontiers in Psychology* **2019**, *10*.
23. Paas, F.G.; Van Merriënboer, J.J. Instructional control of cognitive load in the training of complex cognitive tasks. *Educational psychology review* **1994**, *6*, 351–371.
24. Paas, F.; Tuovinen, J.E.; Tabbers, H.; Van Gerven, P.W. Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist* **2003**, *38*, 63–71.
25. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology* **1988**, *52*, 139–183. doi:https://doi.org/10.1016/S0166-4115(08)62386-9.

26. DeLeeuw, K.E.; Mayer, R.E. A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of educational psychology* **2008**, *100*, 223.
27. Myrtek, M.; Deutschmann-Janicke, E.; Strohmaier, H.; Zimmermann, W.; Lawerenz, S.; Brügner, G.; Müller, W. Physical, mental, emotional, and subjective workload components in train drivers. *Ergonomics* **1994**, *37*, 1195–1203.
28. Wientjes, C.J.; Grossman, P.; Gaillard, A.W. Influence of drive and timing mechanisms on breathing pattern and ventilation during mental task performance. *Biological Psychology* **1998**, *49*, 53–70.
29. Haapalainen, E.; Kim, S.; Forlizzi, J.F.; Dey, A.K. Psycho-physiological Measures for Assessing Cognitive Load. Proceedings of the 12th ACM International Conference on Ubiquitous Computing; ACM: New York, NY, USA, 2010; UbiComp '10, pp. 301–310. doi:10.1145/1864349.1864395.
30. Sirevaag, E.J.; Kramer, A.F.; REISWEBER, C.D.W.M.; STRAYER, D.L.; GRENELL, J.F. Assessment of pilot performance and mental workload in rotary wing aircraft. *Ergonomics* **1993**, *36*, 1121–1140.
31. Wilson, G.F. Air-to-ground training missions: a psychophysiological workload analysis. *Ergonomics* **1993**, *36*, 1071–1087.
32. Rajan, R.; Selker, T.; Lane, I. Task Load Estimation and Mediation Using Psycho-physiological Measures. ACM International Conference on Intelligent User Interfaces; , 2016.
33. French, J.W.; Ekstrom, R.B.; Price, L.A. *Manual for Kit of Reference Tests for Cognitive Factors: John W. French; Ruth B. Ekstrom; Leighton A. Price*; Educational Testing Service, 1969.
34. Züger, M.; Fritz, T. Interruptibility of software developers and its prediction using psycho-physiological sensors. ACM CHI; , 2015.
35. Matkovič, T.; Pejović, V. Wi-mind: Wireless mental effort inference. Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, 2018, pp. 1241–1249.
36. Setz, C.; Arnrich, B.; Schumm, J.; La Marca, R.; Tröster, G.; Ehlert, U. Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on information technology in biomedicine* **2009**, *14*, 410–417.
37. K. Novak, K. Stojmenova, G.J.; Sodnik, J. Assessment of cognitive load through biometric monitoring; Society for Information Systems and Computer Networks: Ljubljana, Slovenia, 2017.
38. Schaule, F.; Johanssen, J.O.; Bruegge, B.; Loftness, V. Employing Consumer Wearables to Detect Office Workers' Cognitive Load for Interruption Management. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT)* **2018**, *2*, 32:1–32:20.
39. Subramanian, R.; Wache, J.; Abadi, M.K.; Vieriu, R.L.; Winkler, S.; Sebe, N. ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing* **2016**, *9*, 147–160.
40. Correa, J.A.M.; Abadi, M.K.; Sebe, N.; Patras, I. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing* **2018**.
41. Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* **2011**, *3*, 18–31.
42. Bilakhia, S.; Petridis, S.; Nijholt, A.; Pantic, M. The MAHNOB Mimicry Database: A database of naturalistic human interactions. *Pattern recognition letters* **2015**, *66*, 52–61.
43. Abadi, M.K.; Subramanian, R.; Kia, S.M.; Avesani, P.; Patras, I.; Sebe, N. DECAF: MEG-based multimodal database for decoding affective physiological responses. *IEEE Transactions on Affective Computing* **2015**, *6*, 209–222.
44. Schneegass, S.; Pfleging, B.; Broy, N.; Heinrich, F.; Schmidt, A. A data set of real world driving to assess driver workload. Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications, 2013, pp. 150–157.
45. Healey, J.A.; Picard, R.W. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems* **2005**, *6*, 156–166.
46. Pavlidis, I.; Dcosta, M.; Taamneh, S.; Manser, M.; Ferris, T.; Wunderlich, R.; Akleman, E.; Tsiamyrtzis, P. Dissecting driver behaviors under cognitive, emotional, sensorimotor, and mixed stressors. *Scientific reports* **2016**, *6*, 25651.

47. Gjoreski, M.; Gams, M.; Luštrek, M.; Genc, P.; Garbas, J.U.; Hassan, T. Machine Learning and End-to-end Deep Learning for Monitoring Driver Distractions from Physiological and Visual Signals. *IEEE Access* **2020**.
48. Schmidt, P.; Reiss, A.; Duerichen, R.; Marberger, C.; Van Laerhoven, K. Introducing wesad, a multimodal dataset for wearable stress and affect detection. Proceedings of the 20th ACM International Conference on Multimodal Interaction, 2018, pp. 400–408.
49. Birjandtalab, J.; Cogan, D.; Pouyan, M.B.; Nourani, M. A non-EEG biosignals dataset for assessment and visualization of neurological status. 2016 IEEE International Workshop on Signal Processing Systems (SiPS). IEEE, 2016, pp. 110–114.
50. Kirschbaum, C.; Pirke, K.M.; Hellhammer, D.H. The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology* **1993**, *28*, 76–81.
51. Bibbey, A.; Carroll, D.; Roseboom, T.R.; Phillips, A.C.; de Rooij, S.R. Personality and physiological reactions to acute psychological stress. *International Journal of Psychophysiology* **2013**, *90*, 28–36.
52. Childs, E.; White, T.L.; de Wit, H. Personality traits modulate emotional and physiological responses to stress. *Behavioural Pharmacology* **2014**, *25*, 493–502.
53. Dixon, K.E.; Thorn, B.E.; Ward, L.C. An evaluation of sex differences in psychological and physiological responses to experimentally-induced pain: a path analytic description. *Pain* **2004**, *112*, 188–196.
54. Wang, P.; Baker, L.A.; Gao, Y.; Raine, A.; Lozano, D.I. Psychopathic Traits and Physiological Responses to Aversive Stimuli in Children Aged 9–11 Years. *Journal of Abnormal Child Psychology* **2012**, *40*, 759–769.
55. Jorgensen, R.S.; Johnson, B.T.; Kolodziej, M.E.; Schreer, G.E. Elevated blood pressure and personality: A meta-analytic review. *Psychological Bulletin* **1996**, *120*, 293–320.
56. Peters, M.L.; Godaert, G.L.R.; Ballieux, R.E.; Heijnen, C.J. Moderation of physiological stress responses by personality traits and daily hassles: less flexibility of immune system responses. *Biological Psychology* **2003**, *65*, 21–48.
57. Cai, R.; Guo, A.; Ma, J.; Huang, R.; Yu, R.; Yang, C. Correlation Analyses Between Personality Traits and Personal Behaviors Under Specific Emotion States Using Physiological Data from Wearable Devices. 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech), 2018, pp. 46–53.
58. Cai, R.; Guo, A.; Ma, J.; Huang, R. Correlational Analyses among Personality Traits, Emotional Responses andBehavioral States Using Physiological Data from Wearable Sensors. eTELEMED 2018 : The Tenth International Conference on eHealth, Telemedicine, and Social Medicine, 2018, pp. 43–46.
59. Sano, A.; Phillips, A.J.; Yu, A.Z.; McHill, A.W.; Taylor, S.; Jaques, N.; Czeisler, C.A.; Klerman, E.B.; Picard, R.W. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. 2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN), 2015, pp. 1–6.
60. Schmiedek, F.; Lövdén, M.; Lindenberger, U. A task is a task is a task: putting complex span, n-back, and other working memory indicators in psychometric context. *Frontiers in psychology* **2014**, *5*.
61. Ashton, M.; Lee, K.; Perugini, M.; Szarota, P.; de Vries, R.; Di Blas, L.; Boies, K.; De Raad, B. A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology* **2004**, *86*, 356 – 366. Article, doi:10.1037/0022-3514.86.2.356.
62. Lee, K.; Ashton, M.C. Psychometric Properties of the HEXACO-100. *Assessment* **2016**, *25*, 543–556. doi:10.1002/1097-4679(198201)38:13.0.CO;2-I.
63. Gravetter, F.J.; Wallnau, L.B. *Essentials of statistics for the behavioral sciences (8th ed.)*; Cengage Learning, 2013.
64. Shaffer, F.; McCraty, R.; Zerr, C.L. A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability. *Frontiers in psychology* **2014**, *5*, 1040.
65. Hoshi, R.A.; Pastre, C.M.; Vanderlei, L.C.M.; Godoy, M.F. Poincaré plot indexes of heart rate variability: relationships with other nonlinear variables. *Autonomic Neuroscience* **2013**, *177*, 271–274.
66. Wang, R.; Tang, K. Feature selection for maximizing the area under the ROC curve. 2009 IEEE International Conference on Data Mining Workshops. IEEE, 2009, pp. 400–405.
67. Arndt, C. *Information measures: information and its description in science and engineering*; Springer Science & Business Media, 2001.

68. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Physical review E* **2004**, *69*, 066138.
69. Ho, T.K. Random decision forests. Proceedings of 3rd international conference on document analysis and recognition. IEEE, 1995, Vol. 1, pp. 278–282.
70. Quinlan, J.R. Improved use of continuous attributes in C4. 5. *Journal of artificial intelligence research* **1996**, *4*, 77–90.
71. Russell, S.J.; Norvig, P. *Artificial intelligence: a modern approach*; Malaysia; Pearson Education Limited, 2016.
72. Aha, D.W.; Kibler, D.; Albert, M.K. Instance-based learning algorithms. *Machine learning* **1991**, *6*, 37–66.
73. Lin, C.J.; Weng, R.C.; Keerthi, S.S. Trust region newton methods for large-scale logistic regression. Proceedings of the 24th international conference on Machine learning, 2007, pp. 561–568.
74. Breiman, L. Bagging predictors. *Machine learning* **1996**, *24*, 123–140.
75. Rosenblatt, F. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Laboratory, 1961.
76. Caruana, R. Multitask learning. *Machine learning* **1997**, *28*, 41–75.
77. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. Proceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 807–814.
78. Rissler, R.; Nadj, M.; Li, M.X.; Knierim, M.T.; Maedche, A. Got flow? Using machine learning on physiological data to classify flow. Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, 2018, pp. 1–6.
79. Maier, M.; Marouane, C.; Elsner, D. Deepflow: detecting optimal user experience from physiological data using deep neural networks. Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 2108–2110.
80. Gjoreski, M.; Luštrek, M.; Pejović, V. My Watch Says I'm Busy: Inferring Cognitive Load with Low-Cost Wearables. Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, 2018, pp. 1234–1240.
81. Lohani, M.; Payne, B.R.; Strayer, D.L. A review of psychophysiological measures to assess cognitive states in real-world driving. *Frontiers in human neuroscience* **2019**, *13*.
82. Mohanavelu, K.; Poonguzhali, S.; Ravi, D.; Singh, P.K.; Mahajabin, M.; Ramachandran, K.; Singh, U.K.; Jayaraman, S. Cognitive Workload Analysis of Fighter Aircraft Pilots in Flight Simulator Environment. *Defence Science Journal* **2020**, *70*, 131.
83. Johannessen, E.; Szulewski, A.; Radulovic, N.; White, M.; Braund, H.; Howes, D.; Rodenburg, D.; Davies, C. Psychophysiologic measures of cognitive load in physician team leaders during trauma resuscitation. *Computers in Human Behavior* **2020**, p. 106393.
84. Fritz, T.; Begel, A.; Müller, S.C.; Yigit-Elliott, S.; Züger, M. Using psycho-physiological measures to assess task difficulty in software development. Proceedings of the 36th international conference on software engineering, 2014, pp. 402–413.
85. Jimenez-Molina, A.; Retamal, C.; Lira, H. Using psychophysiological sensors to assess mental workload during web browsing. *Sensors* **2018**, *18*, 458.
86. Chen, S.; Epps, J. Atomic Head Movement Analysis for Wearable Four-Dimensional Task Load Recognition. *IEEE journal of biomedical and health informatics* **2019**, *23*, 2464–2474.
87. Dearing, D.; Novstrup, A.; Goan, T. Assessing workload in human-machine teams from psychophysiological data with sparse ground truth. International Symposium on Human Mental Workload: Models and Applications. Springer, 2018, pp. 13–22.
88. Wu, Y.; Miwa, T.; Uchida, M. Using physiological signals to measure operator's mental workload in shipping–an engine room simulator study. *Journal of Marine Engineering & Technology* **2017**, *16*, 61–69.
89. Zhao, S.; Ding, G.; Han, J.; Gao, Y. Personality-Aware Personalized Emotion Recognition from Physiological Signals. IJCAI, 2018, pp. 1660–1667.
90. Reynolds, W. Development of reliable and valid short forms of the Marlow–Crowne Social Desirability Scale. *Journal of Clinical Psychology* **1982**, *38*, 119–125. doi:10.1002/1097-4679(198201)38:13.0.CO;2-I.