Detecting Users' Cognitive Load by Galvanic Skin Response with Affective Interference

Article *in* The ACM Transactions on Interactive Intelligent Systems · September 2017

DOI: 10.1145/2960413

4 authors, including:

Fang Chen
Jinan University (Guangzhou, China)
281 PUBLICATIONS **4,368** CITATIONS

SEE PROFILE

Rafael A Calvo
Imperial College London
390 PUBLICATIONS **11,788** CITATIONS

SEE PROFILE

Detecting Users' Cognitive Load by Galvanic Skin Response with Affective Interference

NARGESS NOURBAKHSH, University of Sydney and NICTA, Australia
FANG CHEN, NICTA, Australia
YANG WANG, NICTA, Australia
RAFAEL A. CALVO, University of Sydney, Australia

Experiencing high cognitive load during complex and demanding tasks results in performance reduction, stress, and errors. However, these could be prevented by a system capable of constantly monitoring users' cognitive load fluctuations and adjusting its interactions accordingly. Physiological data and behaviours have been found to be suitable measures of cognitive load and are now available in many consumer devices. An advantage of these measures over subjective and performance-based methods is that they are captured in real-time and implicitly while the user interacts with the system, which makes them suitable for real-world applications. On the other hand, emotion interference can change physiological responses and make accurate cognitive load measurement more challenging. In this work, we have studied six galvanic skin response features in detection of four cognitive load levels with the interference of emotions. The data was derived from two arithmetic experiments and emotions were induced by displaying pleasant and unpleasant pictures in the background. Two types of classifiers were applied to detect cognitive load levels. Results from both studies indicate that the features explored can detect four and two cognitive load levels with high accuracy even under emotional changes. More specifically, rise duration and accumulative GSR are the common best features in all situations, having the highest accuracy especially in the presence of emotions.

Categories and Subject Descriptors: H.1.2 [Models and Principles]: User/Machine Systems---Human factors, Human information processing; H.5.2 [Information Interfaces and Presentation]: User Interfaces---Evaluation/Methodology, User-centered design

General Terms: Human Factors, Measurement

Additional Key Words and Phrases: Cognitive Load, Physiological Data, Galvanic Skin Response, Emotion Interference, Machine Learning

ACM Reference format:

Nargess Nourbakhsh, Fang Chen, Yang Wang, and Rafael A. Calvo, 2017. Detecting Users' Cognitive Load by Galvanic Skin Response with Affective Interference. *ACM Trans. Interact. Intell. Syst.* Vol. xx, No. x, Article xx (Month 2017), 21 pages.

DOI: 10.1145/1234

Authors' addresses: N. Nourbakhsh, School of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia; F. Chen and Y. Wang, NICTA, Level 5, 13 Garden Street, Eveleigh, NSW 2015, Australia; R. A. Calvo, School of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia.

DOI: 10.1145/1234

XX

1 INTRODUCTION

Behaviours (such as gestures and movements) and physiological data (signals from brain, heart, muscles, skin, eyes, etc.) are non-verbal outputs of human body which very often reveal valuable and reliable information about what people think and feel. Modern equipment makes it possible to collect large amounts of this data in real-time, accurately and while people are focusing on their own tasks, without interrupting their normal routine.

With the help of computational procedures and machine learning techniques we can extract meaningful patterns from behavioural and physiological data and relate them to different states of emotions, engagement, cognition, etc. Therefore, intelligent interactive systems can be developed to constantly measure and monitor the current mental status of their users and adjust their interactions accordingly. Such situation-aware personalization is useful in many applications, from improving productivity, preventing mistakes and possibly supporting psychological wellbeing by reducing stress. Common examples include crisis and air-traffic control, traffic management and emergency call handling centers where complex and cognitively demanding tasks may overload a user, potentially leading to errors with serious life-threatening consequences. An adjustable intelligent system can prevent such mistakes by properly distributing the tasks among the users or changing its interface and interactions with them based on their current mental state. These and other examples also highlight the possibility of designing in order to improve psychological wellbeing [5] by reducing stress.

This study contributes to understanding of how meaningful change patterns in human behaviours and physiological signals are related to different levels of cognitive load. More specifically, our recent research has focused on the economical conveniently-captured galvanic skin response (GSR). An important issue for the design of interactive intelligent systems is that emotions, which are often not relevant to the main task, can interfere with the normal process of the related information in the working memory, inducing extra mental load as well as changing physiological data and making the detection of experienced cognitive load even more challenging. In particular this paper investigates the usefulness of galvanic skin response for automatic detection of different levels of cognitive load in the presence and absence of affective stimuli.

1.1 Cognitive Load Measurement

Cognitive load has been defined as the mental effort or the mental load imposed on working memory. It has also been described as a multidimensional construct consisting of causal factors including the characteristics of task, subject, and interactions between them, and assessment factors including *performance*, *mental load* which is imposed by the task or environment demands, and *mental effort* that reflects the actual cognitive capacity or resources allocated to the task [31]. In this paper cognitive load refers to human's working memory capacity or cognitive resources allocated to demands of performing a particular task. In contrast to sizeable long-term memory, working (or short-term) memory is well known to have limited capacity and duration [7]. This means that only a small number of information 'chunks' can be simultaneously kept in the working memory and for a short period of time [9]. Furthermore, humans often use the information held in working memory in some sort of processing (comparing, associating, computing, etc.). Interactions between information elements require additional working memory capacity [42], reducing the available cognitive resources. Overloading the working memory often leads to performance reduction and errors which in some cases (such as air traffic control and

military operations) may have serious consequences. Therefore it is quite necessary to monitor and measure the cognitive load experienced.

The concept of cognitive load (CL) was introduced by educational scientists by the end of the 1980s to describe and facilitate the learning procedure [40]. Since then, researchers have studied ways to improve teaching materials and methods to achieve better learning outcomes based on the architecture of human mental resources and the mechanism of understanding and learning [6, 30, 36, 41, 43]. But education is not the only domain that cognitive load theory (CLT) can help. Mental overload directly affects the safety in aviation, driving and military operations [10, 12, 18] and must be avoided as much as possible. CLT can be effectively used for flight safety by monitoring the mental workload on pilots and air traffic controllers. User interface design can also benefit significantly from CLT, since information representation can have a dramatic effect on the quality of user experience in communication with computer systems [15, 35]. Performance of brain-computer interfacing (BCI) can be vastly improved by monitoring user's current cognitive status [25].

Many empirical methods exist for quantifying cognitive load; however, they can be categorized into four classes of techniques. *Subjective* cognitive load measurement requires that people rate the amount of mental effort or load they have experienced in order to complete a task. Subjective ratings can be numeric (e.g. from 1 to 9 [30]) or verbal (e.g. from 'strongly disagree' to 'strongly agree' [35]), single-dimensional or multi-dimensional (e.g. NASA-TLX [4, 14]). *Performance-based* approaches measure cognitive load by monitoring the achievements of the subjects, such as completion time and the rate of correct answers, in performing an activity [6]. It is also possible to assess the performance on a secondary (dual) task as an indicator of the cognitive load associated with a primary task [22]. Subjective and performance-based measurement techniques have been widely used and, regarding implementation, are usually the most convenient methods. However, asking subjects to rate the experienced mental workload means several interruptions and distractions from performing the principal tasks. Moreover, both methods are post-task processing and can be done when the task is finished, thus are not useful for real-time cognitive load assessment. In contrast, human *behaviours* and *physiological responses* can continuously and non-intrusively demonstrate user's cognitive states while performing the intended task. Speech [18, 46], pen input [34, 47], and eye movements [8, 10] are examples of behaviours that have been studied for mental load assessment. Finally, several physiological signals have been used for cognitive load measurement, including but not limited to, signals from heart [32], eye [45], brain [2], muscles [22] and skin [29, 38, 44]. Since the study of galvanic skin response in cognitive load measurement is the focus of this paper, we first briefly explain this signal and then review some of the previous related studies before continuing onto our recent work.

1.2 GSR: Terminology, Physiology and Measurement

Galvanic skin response (GSR) has been recorded and investigated in thousands of psychophysiological studies. This popularity is partly due to the sensitivity of this signal to the changes in human mental status and processes, and on the other hand to the fact that it can be quantified in relatively convenient and low-cost manners which makes it suitable not only in the laboratory experiments but also in real world applications.

The term to address the electrical activity of the skin has changed over the years. In the nineteen century it was called *psychogalvanic reflex* [39]. Later, the term *Electrodermal activity*

(EDA) was introduced to use for all electrical phenomena of the skin [3, 20]. Nowadays the terms EDA, GSR and skin conductance response (SCR) are alternatively used.

The electrical activity of the skin can be measured in two ways. In *exosomatic* methods, a small current - which can be either alternating current (AC) or direct current (DC) - from an external source is passed through the skin and the resistance of the skin to the passage of the current is measured. *Endosomatic* methods on the other hand, measure the potential differences at the surface of the skin without using any external current. Today, most researchers apply the exosomatic method in which the *skin conductance* (reciprocal of resistance) is measured [3, 39].

It is commonly agreed that *eccrine* sweat glands (a special type of sweat glands) are closely involved in producing GSR [3, 11, 39]. Eccrine sweat glands are mostly found in the palms of hands and soles of feet and are controlled by the sympathetic branch of the autonomous nervous system (ANS). Interestingly, secretion of this type of sweat glands is found to be related to the psychological and processing stimuli, while other sweat glands mostly act in temperature regulation [39].

Based on the impulses from the sympathetic system, the number of active sweat glands, the level of sweat in each gland and the amount of sweat which overflows on the skin surface vary. Sweat ducts (small tubes which open from the secretion segment of the gland to the skin surface) act as tiny variable resistors in parallel and the total conductance of a parallel circuit is simply the sum of all the conductance of the active resistors. This is one of the reasons why using skin conductance is preferred to using skin resistance in exosomatic method. Further information about the anatomical origin, physiology and recording of GSR can be found in [1, 3, 11].

1.3 GSR in Detecting Cognitive Load and Emotions

GSR, individually or with other physiological signals and data, has been investigated concerning different aspects of human mind in various experimental scenarios. Reviewing all the previous work in this area is beyond the limitations of the present paper, thus we summarize only some of the key and recent studies on using GSR in detection of human emotions and mental workload.

Nakasone et al. have successfully used skin conductance and muscle activity for emotion detection in a computer card game [27]. In another study, skin conductance was measured to differentiate between a stress condition and a cognitive load condition in an office environment, seeking the ability of detecting stress states [37]. Shi et al. also assessed GSR in multiple stress and cognitive load situations in a traffic management application and found correlations between mean and summation of this signal and cognitive load [38]. Ikehara and Crosby [19] evaluated GSR together with body temperature, blood flow, blood oxygen and pulse rate, eye movement and mouse pressure in relation with two levels of cognitive load in a moving targets game. In contrast with other studies, they found skin conductance to decrease as task difficulty increases. They explained this as a result of the easy task being tedious and too easy.

Investigating the effect of cognitive and visual demands on driving performance and driver's status, Engstrom et al. found a weak effect of cognitive load on physiological signals including mean skin conductance [12]. In another on-road study Mehler et al. found skin conductance and heart rate to increase with higher cognitive demand across three difficulty levels [24]. They suggest that Engstrom et al.'s less consistent results might be partially caused by participant disengagement. Wilson [44] analysed several physiological measures during different steps of flights and found an increase in EDA response during take-off and landing which are expected to place the most cognitive demands on pilots. Haapalainen, Kim, Forlizzi and Dey [13] assessed

mean, variance and median of GSR and several other physiological signals against two cognitive load levels in a set of small visual attention demanding tests (consisting of patterns, words, and numbers). They did not obtain any satisfactory results for GSR and explained that it might be related to the tasks type or their GSR sensors might not have been sensitive enough [13]. Some studies have been conducted in which GSR features, rather than being separately evaluated, comprise a part of a multichannel physiological feature and the performance of that single physiological modality is assessed in emotion or cognitive load detection [16, 17].

Nourbakhsh et al. evaluated temporal and spectral features of GSR in reading and arithmetic tasks consisting of three and four difficulty levels respectively and found strong correlations between the studied GSR features and cognitive load levels [29]. In another study on arithmetic tasks with four difficulty levels, GSR and eye blink features were found to classify cognitive load levels with reasonable accuracy and combination between the two modalities improved the classification results [28]. In this paper, we investigate the performance of six GSR features in detecting cognitive load levels in two different arithmetic experiments. There have been many differences in the design, recording devices and protocols, and participants of the two studies. On the other hand, both experiments consisted of four difficulty levels and a task-only part as well as interference of emotions. Two types of machine learning algorithms have been applied for four-class and binary classification of cognitive load in each experiment. We have assessed the statistical and classification behaviour of the features in the presence and absence of emotions and compared them between the two experiments.

2 FIRST STUDY

2.1 Experiment Design

This experiment consisted of three parts. In the first part, the background was black and only the numbers were displayed on the screen. In the second and third part, pleasant and unpleasant images were displayed in the background while the subject was performing the tasks. Therefore, the first part was task-only part whereas in the rest of the experiment tasks were performed under the interference of emotions. The pictures were selected from the International Affective Picture System (IAPS) database which contains color images with normative emotional stimuli and is widely used in studies on emotion and attention [21]. Table 1 shows average normative ratings and examples of the images used to induce emotions in this experiment.

Table 1. Average normative ratings and examples of IAPS images used in first experiment

Category	Mean Valence	Mean Arousal	Examples
Pleasant (Positive)	7.1	5.7	children, romance, gold
Unpleasant (Negative)	2.8	4.8	snakes, skulls, injury

The whole experiment included 24 arithmetic tasks with 4 difficulty levels. In each of the three parts, subjects performed two trials of each task level and the eight tasks of each part were presented in a randomized order. First to fourth difficulty levels respectively included binary numbers (0 and 1), one-digit numbers (1 to 9), two-digit numbers (10 to 99) and three-digit

numbers (100 to 999). In each task four numbers were shown one by one, each for three seconds. Subjects were supposed to add-up these four numbers and select (by clicking the mouse using their right hand) the correct answer from three numbers which were next presented on the screen. Trials with a same difficulty level included different groups of numbers.

Before the first number of each task was displayed, a slide containing one to three 'x' symbols (according to the number of digits in the task) was presented for three seconds. There was no time limit for answer input. A 60-second resting period was recorded before commencing and after finishing the whole experiment while the subjects were asked to relax and look at a black screen. There was a 6-second rest time between consecutive tasks. For the second and third part of the experiment, the image was displayed in this 6-second period as well as the following task. The data collection of this experiment took about 15 minutes for each participant. After the experiment, subjects were asked to rate the difficulty level of each task in a 9-point questionnaire. The questionnaire was displayed on the screen, the scores ranged from 1 (for extremely easy) to 9 (for extremely hard) and the participants selected the scores by mouse click.

2.2 Apparatus

To collect galvanic skin response, the GSR device from ProComp Infiniti of Thought Technology Ltd was used and the sensors were attached to the subject's index and ring fingers of left hand. The sensors were placed in contact with the outer segment of each finger. The sampling frequency was 10Hz. Eye activity data was also recorded with a remote eye tracker (faceLAB 4.5 of Seeing Machines Ltd) which operated at a sampling rate of 50Hz. A 21" LCD Dell monitor and a conventional computer mouse were peripherals for interaction between participants and a PC running the tasks. Another PC collected the signals through GSR sensors and eye tracker and was synchronized with the first one. This paper focuses on investigation and findings related to GSR across the two studies, and research and results about eye tracking data have been reported in other publications.

2.3 Participants

Twelve healthy 21 to 35-year-old (M=30, SD=8.04) male volunteers (students and staff members) participated in this experiment. All subjects were right-handed. They signed an informed consent form before the experiment and were rewarded with movie vouchers for their participation. The experiment was approved by Human Research Ethics Committee of the University of New South Wales.

2.4 Data Analysis

2.4.1. Feature Extraction. For each task, GSR features were extracted from the part of signal corresponding to task time by which we mean from presenting the first number until clicking on the answer. Each task on average lasted between 13.284 and 14.608 seconds depending on the difficulty level. Six GSR features were calculated for each task:

- Peak number: number of total peaks (responses) in the task.
- Peak amplitude: sum of response amplitudes of all responses over the task time. Response amplitude is the difference of GSR values between the point the signal starts to rise (response onset) until it reaches to its local maximum (response peak).

- Rise duration: sum of rise durations of all responses over the task time. Rise duration is the interval between response onset and response peak.
- Peak area: Sum of response areas (0.5 × amplitude × duration).
- Accumulative GSR: the summation of signal values over task time.
- Frequency power: power spectrum of the signal (calculated using the Fast Fourier Transform) over task time; the DC part was discarded.

Figure 1 illustrates a part of GSR signal from one of the subjects while they were performing the arithmetic tasks, and also illustrates how we calculated peak amplitude and rise duration.

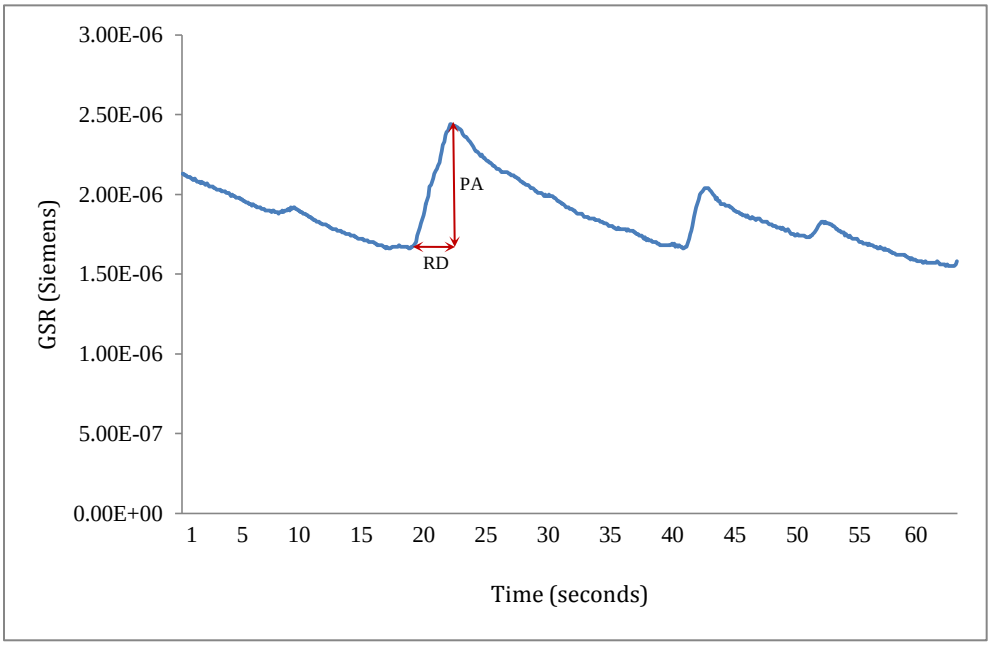

Figure 1. Sample GSR signal from the first experiment, peak amplitude (PA) and rise duration (RD).

2.4.2. Feature Normalisation. GSR signal is highly individual dependent, in other words the average skin conductance varies widely between individuals. To omit this subject dependency, we have calibrated each feature of every subject, dividing its value by the mean value of all the same features of that specific subject over all tasks.

2.4.3. Statistical Analysis. To assess the statistical significance, we have applied one-way analysis of variance (ANOVA test, α=0.05) on the subjective ratings as well as GSR features. For each subject and each feature, the feature values for similar difficulty levels have been averaged. We have also calculated the effect size (using Cohen's d) to compare the effectiveness of different features and performed paired t-test to compare the difference between pairs of cognitive load levels for each feature. Some researchers recommend post-hoc adjustment, i.e. applying a more rigorous significance level, when performing more than one statistical test. Others on the other hand believe that such adjustments are unnecessary or even wrong [26, 33]. Due to this controversy we performed pair-wise analysis twice, first without post-hoc adjustment (significance

level=0.05) and then after post-hoc adjustment using Bonferroni correction (significance level=0.0083).

2.4.4. Classification. We have classified the GSR features using support vector machines (SVM) and Naïve Bayes classifiers. These supervised machine learning algorithms are among the popular ones in human-computer interaction (HCI) studies due to their implementation simplicity and relatively high performance [23]. For each feature and in each experiment, we have examined four-class and binary classification. In binary classification, we have considered levels one and two as low cognitive load and levels three and four as high cognitive load. In order to keep training and testing data independent, we have applied leave-one-subject-out method for cross validation. That is, we have trained the classifiers with data of all subjects but one, and then used the remaining subject data as the testing data. We repeated this procedure for each subject. At the end, we averaged the classification accuracies over all subjects.

Custom code was written for feature extraction and normalisation. MATLAB built-in functions were used within our custom code for classification and statistical analyses. The abovementioned procedures and methods were applied for this study as well as the second study (Section 3). In each study, the entire analysis was once performed on the task-only part of the data and then it was repeated on the whole experiment data (including emotional interference).

2.5 Results

ANOVA test on self-report scores produced significant results ($F(3,44)=108.63$, $p<0.001$) and there was an increasing trend between the subjective ratings and difficulty levels (Figure 2). This means that the experimental design has successfully manipulated the experienced cognitive load levels.

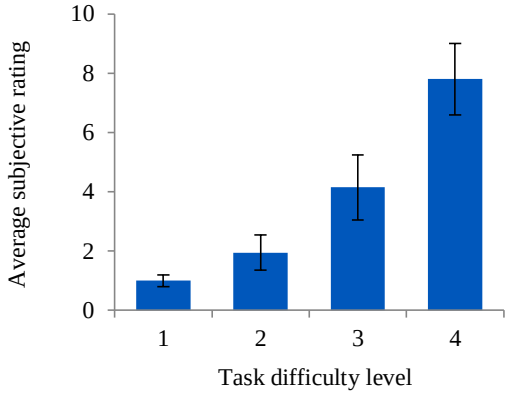

Figure 2. Average subjective ratings of the difficulty levels for the first experiment; error bars show standard deviation (SD).

Table 2 shows the results of ANOVA test and effect sizes of the GSR features in the first (task-only) part of the experiment. ANOVA test results indicate that all the features significantly differentiate between the four cognitive load levels. Comparing the effect sizes shows that rise duration, accumulative and peak number are the most effective features respectively.

Table 2. ANOVA test results and effect sizes of GSR features in the first experiment (task-only part)

Feature	F(3,44)	p-value	Effect Size
Peak number	6.12	0.001	0.243
Peak amplitude	2.90	0.046	0.096
Rise duration	16.36	<0.001	0.490
Peak area	2.92	0.044	0.103
Accumulative	7.57	<0.001	0.291
Frequency power	3.74	0.018	0.146

In order to evaluate the statistical difference between each difficulty level and other levels, paired t-test was performed on pairs of cognitive load levels for each feature. Table 3 shows that each difficulty level has a significant difference from other levels ($p<0.05$, bolded values). The best results relate to the comparison of level 1 with levels 2 and 4. The majority of difficulty level pairs are significantly different for rise duration and accumulative GSR. After Bonferroni adjustment on this dataset ($\alpha=0.0083$), some significant results become insignificant (indicated by *). As discussed in section 2.4.3 there are disagreements about the necessity and validity of such adjustments.

Table 3. Paired t-test results for the GSR features in the first experiment (task-only part)

Feature	CL1 vs. CL2	CL1 vs. CL3	CL1 vs. CL4	CL2 vs. CL3	CL2 vs. CL4	CL3 vs. CL4
Peak number	**p=0.0095***	p=0.2137	p=0.2938	p=0.3121	**p=0.0077**	p=0.0722
Peak amplitude	p=0.1461	p=0.1165	**p=0.0302***	p=0.6034	p=0.2893	p=0.6881
Rise duration	p=0.9603	p=0.0654	**p=0.0011**	**p=0.0475***	**p<0.0001**	**p=0.0192***
Peak area	**p=0.0336***	p=0.1416	**p=0.0455***	p=0.3680	p=0.9711	p=0.4060
Accumulative	**p=0.0243***	**p=0.0298***	**p=0.0053**	p=0.1765	**p=0.0264***	p=0.2048
Frequency power	**p=0.0176***	**p=0.0164***	p=0.2548	p=0.8993	p=0.2719	p=0.2339

Average classification accuracies of the features on the no-emotion part of the experiment are illustrated in Figures 3 and 4. Baselines were 25% and 50% for four-class and binary classification respectively, and same baseline levels were used throughout the paper.

For four-class classification (Figure 3), almost all the features have high accuracies with both classifiers. Peak number and rise duration have highest accuracies (39.6% to 41.7%); slightly lower are peak amplitude, accumulative and frequency power (around 35% to 37.5%). Peak area has the highest difference between the two classifiers: 35.4% with Naïve Bayes and 25% with SVM.

Binary classification results of the first experiment in the absence of emotions are presented in Figure 4. It can be observed that SVM and Naïve Bayes classifiers have almost or exactly similar results for each feature. All features have accuracies higher than baseline, however peak number,

rise duration, accumulative and frequency power (around 70%) are performing better than peak amplitude and peak area (between 54.2% and 60.4%).

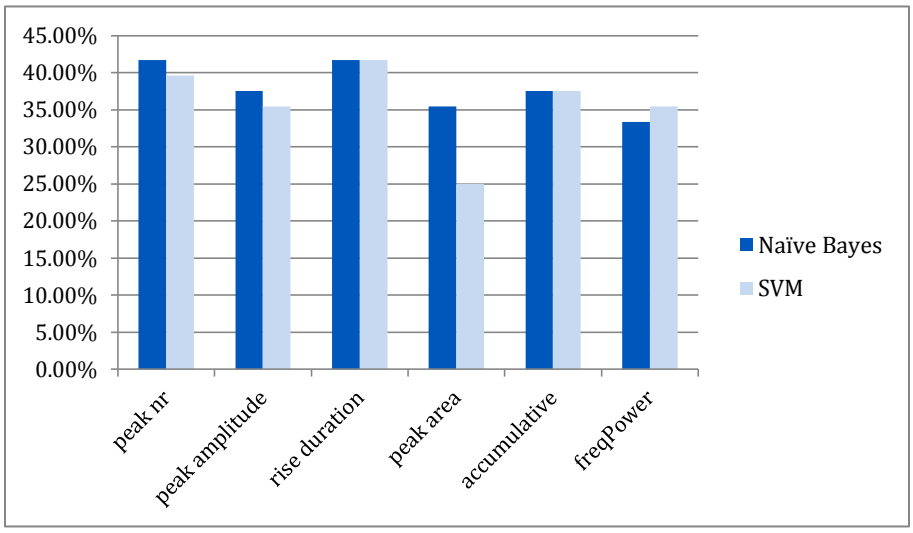

Figure 3. Four-class classification results of first experiment (no-emotion part)

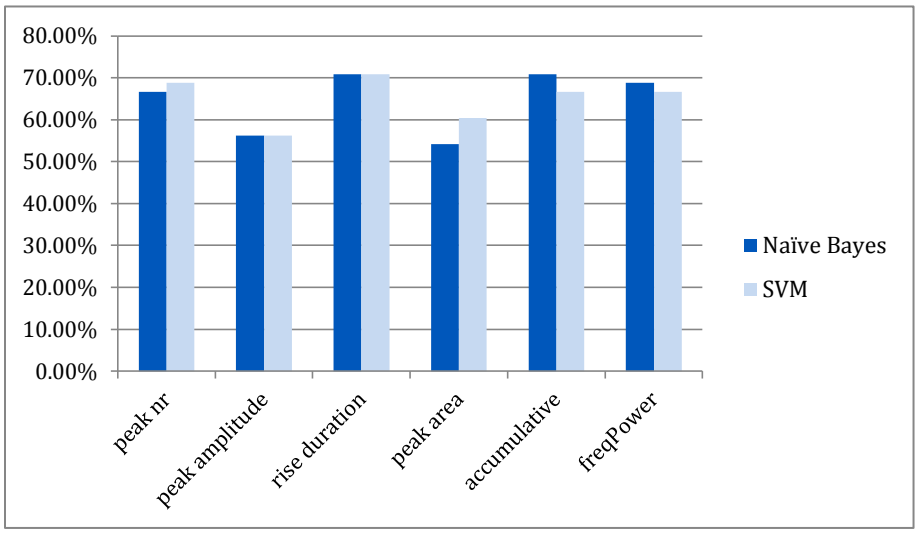

Figure 4. Two-class classification results of first experiment (no-emotion part)

Now we turn to the results of the next part of the analysis on GSR features of our first dataset in which performing of the arithmetic tasks have been interfered by induction of positive and negative emotions. It should be noted that the primary focus of this paper was to investigate the performance of GSR features in cognitive load measurement, and emotions were induced as a

confounding factor to evaluate robustness of the features in that regard. Therefore analysis of the impact of affective stimuli on GSR features was beyond the limitations of this paper and not reported here.

As it can be observed from Table 4, all GSR features have significant statistical results on distinction of the four cognitive load levels in the presence of emotions; however, effect sizes show that accumulative GSR and rise duration are the most effective features.

Table 4. ANOVA test results and effect sizes of the GSR features in the first experiment (including emotional changes)

Feature	F(3,44)	p-value	Effect Size
Peak number	3.18	0.033	0.120
Peak amplitude	5.56	0.003	0.222
Rise duration	14.28	<0.001	0.454
Peak area	3.04	0.039	0.094
Accumulative	21.05	<0.001	0.555
Frequency power	3.26	0.030	0.124

We performed paired t-tests to compare the differentiation between pairs of cognitive load levels for each feature. Results are provided in Table 5 and significant values before post-hoc adjustment (α=0.05) are bolded. Each difficulty level has a significant difference from the other levels. In particular, level 1 is significantly different from levels 3 and 4, and accumulative GSR causes the best results following by rise duration. After post-hoc adjustment (α=0.0083) some of the significant results would be considered as insignificant (indicated by *). Rise duration and accumulative GSR still produce the best results.

Table 5. Paired t-test results for GSR features in the first experiment (including emotional changes)

Feature	CL1 vs. CL2	CL1 vs. CL3	CL1 vs. CL4	CL2 vs. CL3	CL2 vs. CL4	CL3 vs. CL4
Peak number	p=0.1268	p=0.9515	p=0.2155	p=0.1068	p=0.0683	p=0.2627
Peak amplitude	p=0.2581	**p=0.0109***	**p=0.0032**	p=0.2530	p=0.1188	p=0.4383
Rise duration	p=0.2987	**p=0.0069**	**p<0.0001**	p=0.0840	**p=0.0029**	p=0.0770
Peak area	p=0.1588	**p=0.0162***	**p=0.0459***	p=0.8806	p=0.5084	p=0.5265
Accumulative	p=0.0865	**p=0.0006**	**p=0.0005**	**p=0.0065**	**p=0.0035**	**p=0.0295***
Frequency power	**p=0.0349***	**p=0.0121***	p=0.4234	p=0.5749	p=0.2682	p=0.2927

Figure 5 shows the average classification accuracies of GSR features for four cognitive load classes. The highest accuracies are obtained from accumulative (around 50%) and rise duration (around 48%), and then peak number (with SVM), peak amplitude and frequency power (around 40%). Naïve Bayes on peak number and peak area results in the average accuracy of 35.4%, and the lowest accuracy is related to SVM on peak area (29.2%).

Binary classification results are presented in Figure 6. The best results are related to accumulative (83.3% and 85.4%), rise duration (77.1% and 83.3%) and peak amplitude (75% and

77.1%). Around 60% we have peak number and peak area while the latter (just like in Figures 3, 4 and 5) has the highest difference between the performances of the two classifiers. The lowest accuracies belong to frequency power.

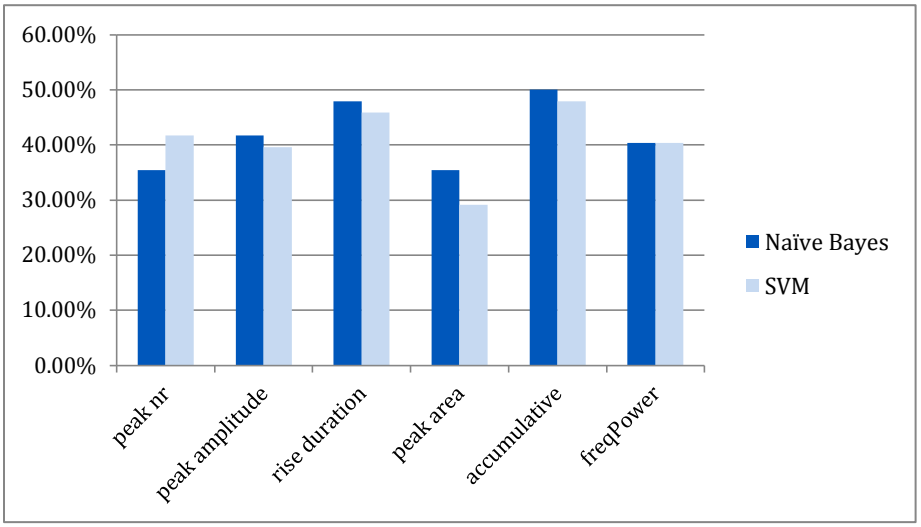

Figure 5. Four-class classification accuracies of first experiment (including emotional changes)

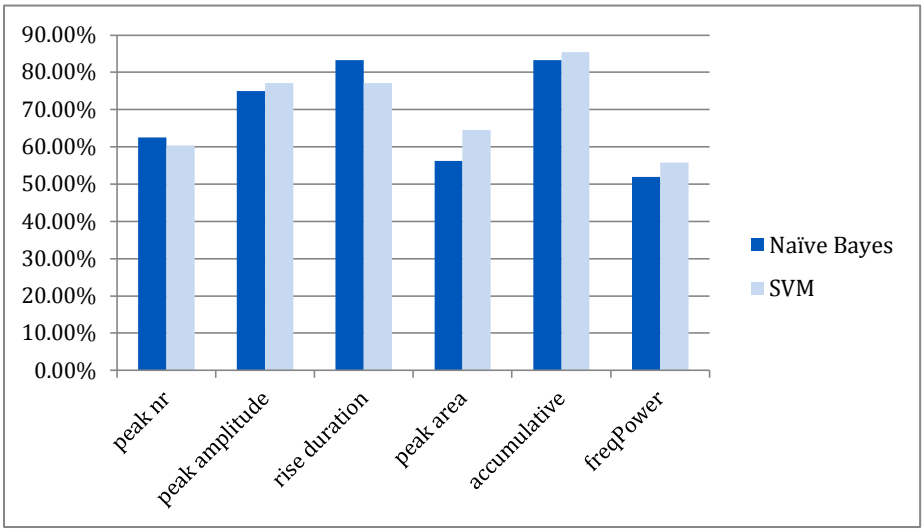

Figure 6. Two-class classification accuracies of first experiment (including emotional changes)

3 SECOND STUDY

3.1 Experiment Design

First study provided positive results about the usability and robustness of GSR features. We tried to improve the experimental design in the second study. As the following will explain, in the

second experiment emotions were controlled more accurately, arithmetic tasks were designed more precisely and difficulty levels were increased more evenly. To increase the reliability and validity of the results, larger dataset was produced (using higher sampling rate and larger number of tasks and participants) and we tried to control some other confounding factors, for example asked participants to avoid caffeine consumption and take no medications prior to the experiment. Parts of the experiment that were similar to the first study are not repeated here.

In total, 56 arithmetic tasks in 7 sessions were performed by each subject. First session was the task-only part and the next 6 sessions used different pictures from 6 categories of valence and arousal levels. Three sessions included positive valence images and three sessions included negative valence pictures. In each of these sessions the image arousal level was high, medium or low. Table 6 shows range of normative ratings and examples of the images used in each session to induce emotions. First and second difficulty levels included one-digit numbers with no carry and one carry respectively produced during the summation. Levels three and four consisted of two-digit numbers; in the former only one carry was produces in the lower digit while in the latter carries were generated in both digits.

Table 6. Range of normative ratings and examples of IAPS images used in second experiment

Category	Valence	Arousal	Example
PositiveValence_LowArousal	>6.42	<3.6	Flower
PositiveValence_MediumArousal	>6.62	>3.94&<5.42	Baby
PositiveValence_HighArousal	>6.21	>5.48	Fireworks
NegativeValence_LowArousal	<4.69	<4.38	Garbage
NegativeValence_MediumArousal	<3.69	>4.54&<5.2	Crying
NegstiveValence_HighArousal	<3.65	>5.79	Snake

At the beginning of each task, ten 'X' or 'XX' symbols were displayed on the grey or image background for 2 seconds which were randomly replaced by the numbers during the task. At the end of the task, ten possible answers were displayed to choose from. Participants were allowed to change the answer as many times as the wished and then submit it. Immediately after each task, a 9-scale questionnaire was displayed to collect the subjective rating of the difficulty level of the task. After the self-report, there was a 5 second resting period with a blank screen and then the next task started. The data collection took about 30 minutes for each participant.

3.2 Apparatus

GSR, together with ECG (electrocardiogram) and respiration data, was recorded using a BIOPAC MP150 system at sampling rate of 1000 Hz. GSR straps were placed around the index and middle fingers of the left hand and a strap to measure respiration was fastened around the chest. ECG was recorded with two sensors attached to subjects' wrists. Eye activity was recorded through a FaceLab 4 remote eye tracker from Seeing Machines and sampling rate was 60 Hz. A Logitech Webcam Pro 9000 was placed on top of the monitor and captured participants' video during the experiment. Peripherals used during this experiment were of the same types as those used in the first study. This paper focuses on the research and findings about GSR features across the two studies; investigation and results related to ECG, respiration and eye tracking data have been reported in other publications.

3.3 Participants

Twenty (11 males and 9 females) healthy students and staff members aged 22-48 (M=27.65, SD=6.94) took part in this experiment. They were asked to avoid caffeine and take no medications before the experiment. The rest of considerations were the same as those applied in the first study.

3.4 Data Analysis

The data processing methods used for this experiment were almost the same as those of the first (described in section 2.4). For each task, the same features were extracted: peak number, peak amplitude, peak duration, peak area, accumulative GSR and frequency power. However, to reduce the computation GSR signal was down-sampled to 200 Hz. Similar to the previous experiment, we have calibrated the features with their subjective average, used one-way ANOVA test with significance level of 0.05, calculated effect sizes using Cohen's d, performed paired t-test at level 0.05 for comparing pairs of difficulty levels and applied SVM and Naïve Bayes classifiers with leave-one-subject-out cross validation method. Custom code was written for feature extraction and normalisation. MATLAB built-in functions were used within our custom code for classification and statistical analyses.

3.5 Results

The average self-report scores were found to increase with the ascending of the difficulty levels (Figure 7). In addition, ANOVA test on the subjective ratings led to significant results ($F(3,72)=79.76$, $p<0.001$). Therefore, this experiment imposed different cognitive load levels on the participants' working memory in the desired order.

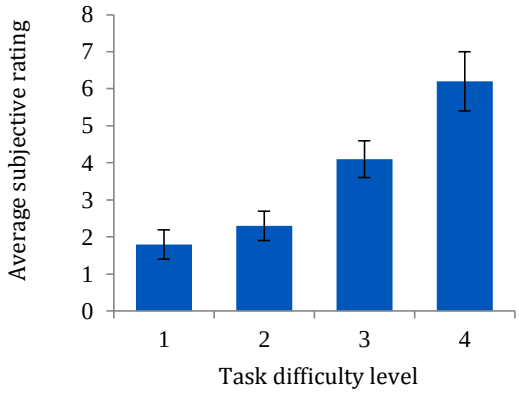

Figure 7. Average subjective ratings of the difficulty levels for the second experiment; error bars show standard deviation (SD).

The results of ANOVA test and effect sizes of the features in detecting four levels of cognitive load without any emotional changes are presented in Table 7. All features significantly distinguish among the four difficulty levels and accumulative GSR and peak rise duration show the highest effectiveness.

We performed paired t-test to compare the differentiation between pairs of cognitive load levels for each feature. Table 8 shows the results and significant ones ($p<0.05$) are bolded. In most cases there is a significant difference between pairs of load levels and each difficulty level has a significant difference from the other levels. After post-hoc adjustment ($\alpha=0.0083$), some results lose significance (indicated by *).

Table 7. ANOVA test results and effect sizes of GSR features of the second experiment (task-only part)

Feature	F(3,72)	p-value	Effect Size
Peak number	9.46	<0.001	0.250
Peak amplitude	17.46	<0.001	0.394
Rise duration	20.31	<0.001	0.433
Peak area	13.19	<0.001	0.325
Accumulative	29.54	<0.001	0.530
Frequency power	9.08	<0.001	0.242

Table 8. Paired t-test results for the GSR features in the second experiment (task-only part)

Feature	CL1 vs. CL2	CL1 vs. CL3	CL1 vs. CL4	CL2 vs. CL3	CL2 vs. CL4	CL3 vs. CL4
Peak number	p=0.8307	p=0.1209	**p=0.0003**	p=0.1053	**p=0.0003**	p=0.0907
Peak amplitude	**p=0.0287***	**p=0.0010**	**p<0.0001**	**p=0.0064**	**p=0.0005**	p=0.1244
Rise duration	**p=0.0293***	**p=0.0025**	**p<0.0001**	**p=0.0217***	**p=0.0004**	**p=0.0245***
Peak area	p=0.1020	**p=0.0029**	**p<0.0001**	**p=0.0178***	**p=0.0012**	p=0.1739
Accumulative	p=0.5902	**p=0.0113***	**p<0.0001**	**p=0.0041**	**p<0.0001**	**p=0.0013**
Frequency power	p=0.6152	**p=0.0166***	**p=0.0087***	**p<0.0001**	**p=0.0132***	p=0.1038

Figure 8 presents the average accuracies of detecting four cognitive load classes when there has been no affective interference. All features have more than 40% accuracy with at least one type of classifiers, peak amplitude, rise duration and accumulative reaching 47.4%, 48.7% and 51.3% respectively. The largest difference between the performance of SVM and Naïve Bayes classifiers relates to accumulative (almost 11.8%) which is followed by peak number (7.9%).

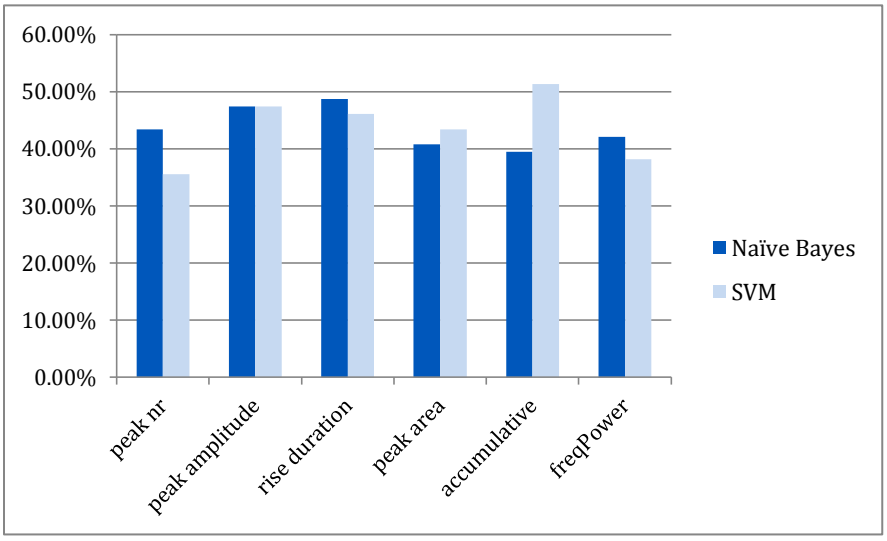

Figure 8. Four-class classification results of the second experiment (non-emotion part)

Binary classification of cognitive load levels on task-only part of the data resulted in 68.4% for peak number, 73.7% for peak area with Naïve Bayes, 75% for peak amplitude and SVM on peak area and frequency power, and around 80% for rise duration, accumulative and Naïve Bayes on frequency power (Figure 9). The frequency feature shows the largest between-classifiers difference (5.3%).

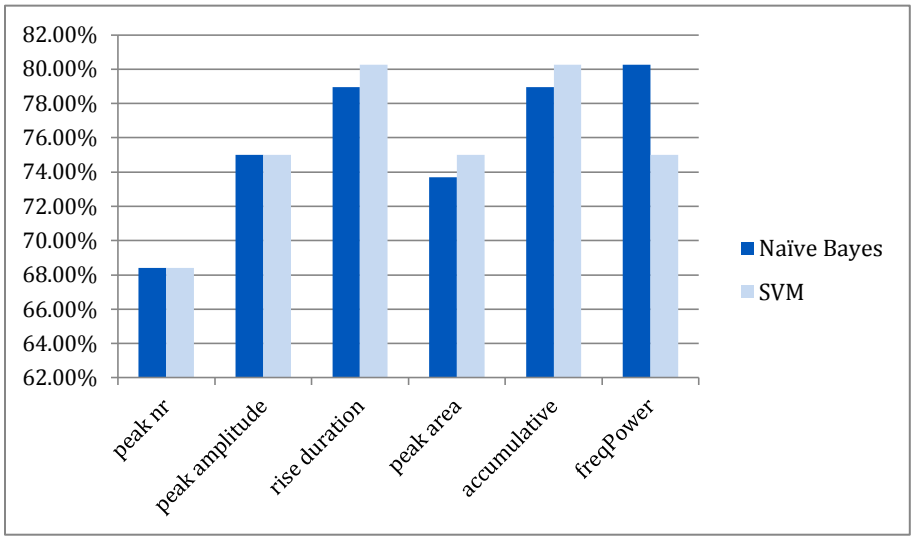

Figure 9. Two-class classification results of the second experiment (non-emotion part)

As the last part of the results, now we describe the outcomes of the statistical analysis and classification of the cognitive load with affective interference in the second experiment. As it can be observed in Table 9, all the features have significant ANOVA results for the four task difficulty levels. Again, rise duration and accumulative GSR are the most effective features in differentiation between cognitive load levels (have the highest effect sizes).

Table 9. ANOVA test results and effect sizes of the GSR features in the second experiment (including emotional changes)

Feature	F(3,72)	p-Value	Effect Size
Peak number	27.56	<0.001	0.512
Peak amplitude	19.27	<0.001	0.419
Rise duration	48.04	<0.001	0.650
Peak area	14.96	<0.001	0.355
Accumulative	42.12	<0.001	0.619
Frequency power	7.87	<0.001	0.213

In order to evaluate the statistical difference between each difficulty level and other levels, paired t-test was performed on pairs of cognitive load levels for each feature. Table 10 shows that

before post-hoc adjustment there is a significant difference between each difficulty level and the other levels. The best results relate to levels 3 and 4, both having significant difference with all other levels for all the features (except the frequency feature for which only level 2 is significantly different). After post-hoc adjustment (α=0.0083) some significant results become insignificant (marked by *).

Table 10. Paired t-test results for GSR features in second experiment (including emotional changes)

Feature	CL1 vs. CL2	CL1 vs. CL3	CL1 vs. CL4	CL2 vs. CL3	CL2 vs. CL4	CL3 vs. CL4
Peak number	p=0.4329	**p=0.0002**	**p<0.0001**	**p=0.0016**	**p<0.0001**	**p=0.0263***
Peak amplitude	p=0.5644	**p=0.0078**	**p<0.0001**	**p=0.0050**	**p<0.0001**	**p=0.0193***
Rise duration	p=0.1936	**p<0.0001**	**p<0.0001**	**p=0.0008**	**p<0.0001**	**p=0.0003**
Peak area	p=0.7657	**p=0.0348***	**p=0.0002**	**p=0.0144***	**p=0.0001**	**p=0.0310***
Accumulative	**p=0.0018**	**p<0.0001**	**p<0.0001**	**p=0.0010**	**p<0.0001**	**p=0.0002**
Frequency power	**p=0.0007**	p=0.8687	p=0.4117	**p=0.0027**	**p=0.0098***	p=0.5282

In four-class cognitive load classification (Figure 10), rise duration has the highest accuracies (57.9% and 55.3%) and the results of accumulative GSR are slightly lower (56.6% and 52.6%). Peak number and amplitude are next (around 47% and 44% respectively). Lowest accuracies as well as the largest difference between the two types of classifiers are related to peak area (36.8% and 43.4%) and frequency power (39.5% and 34.2%).

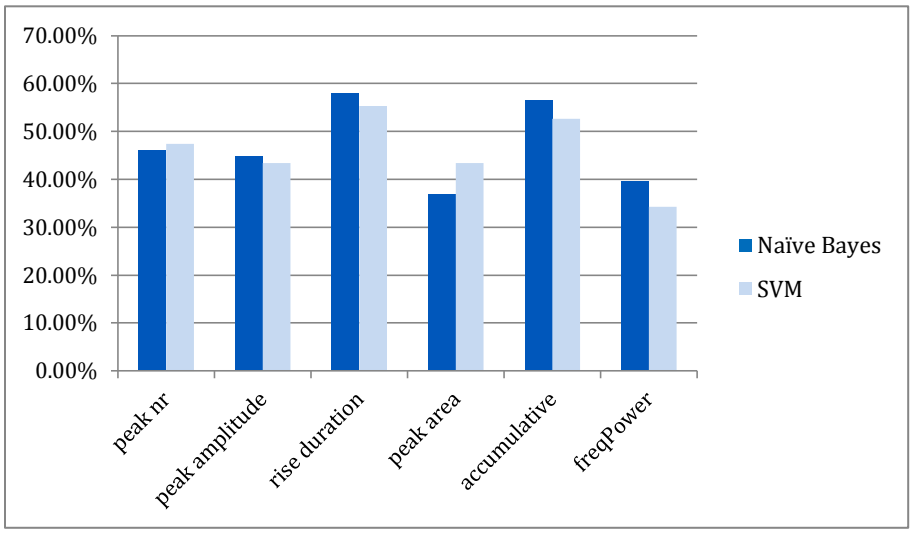

Figure 10. Four-class classification accuracies of the second experiment (including emotional changes)

Figure 11 depicts the binary classification results among which rise duration and peak number have the highest accuracy (86.8%). Then there are accumulative (84.2% and 81.6%) and peak amplitude (77.6% and 80.3%) which is followed by peak area by almost 5% difference. Results of

frequency power, although still above the baseline, are far less than those of the other features (between 20% and 30% below any of them).

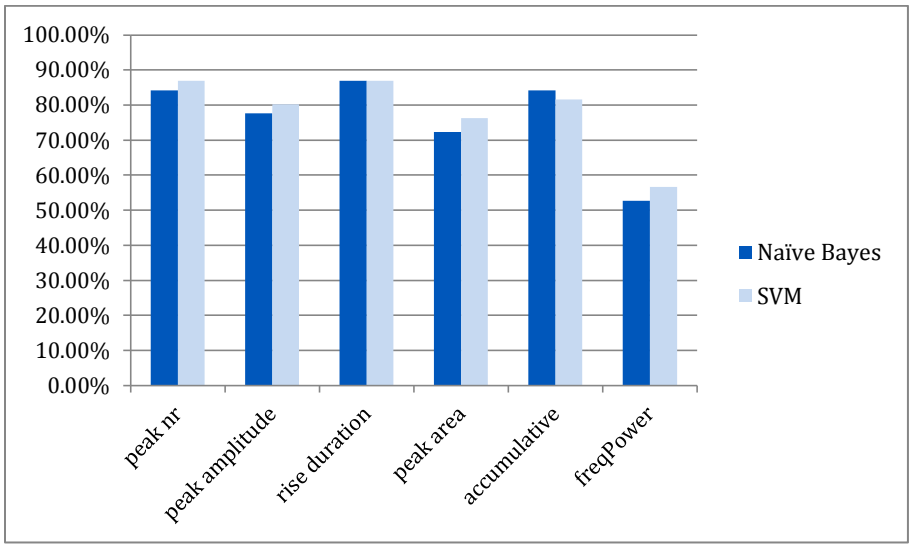

Figure 11. Two-class classification accuracies of the second experiment (including emotional changes)

4 Discussion and Conclusion

The research presented here examined the performance of six GSR features in detecting four levels of cognitive load in two arithmetic experiments. Despite the obvious differences in the design and data recording of the two experiments, the main outcomes of the analysis are shared. All the studied features have significantly differentiated between the four cognitive load levels, in both experiments and regardless of affective interference. The classification accuracy of most of the features in all situations was promising and the two types of classifiers performed closely in most cases. Rise duration and accumulative GSR were the common best features in statistical analysis and classification of two and four cognitive load levels with and without affective interference in the two experiments. For most features results further improved with adding the data with emotional changes. This can be due to the increase in the number of data points.

In general, the second experiment produced better results, both statistically and in classification. Some possible reasons include more accurate recording device, higher sampling rate and larger numbers of tasks and participants (therefore larger amount of data), better task design, and controlling other confounding factors such as caffeine consumption and taking medications. These differences in the design and implementation of the two experiments may also explain the greater range of self-report scores in the first study.

Many of the physiological responses in the human body are controlled by the autonomic nervous system (ANS). This part of the nervous system regulates and coordinates the activities which we almost have no control on and are done without being thought about, such as breathing, circulatory system activity, digestion and sweating. Sympathetic nervous system (SNS) is part of ANS which controls the reflexes necessary to survive in emergency situations, conditions which require alertness or strength, or arise fear, anger or excitement. Higher cognitive demand and emotional arousal stimulate SNS by alerting the body about a challenging situation. SNS activates the proper survival responses such as increased activity of the sweat glands in hands and feet to

increase bodily efficiency [1]. Therefore, fluctuations in cognitive load and emotions lead to changes in the patterns of GSR features.

The relation between GSR and cognitive load had been previously assessed but most studies had consisted of only two cognitive load levels and focused on mean GSR. A few of them had found some correlations [38, 44] and some did not succeed in relating GSR and cognitive load levels [13] or found weak relations [12]. The present research however has found significant statistical and classification results for using temporal and spectral GSR features in detecting two and four cognitive load levels. Most of the features we investigated had not been previously studied. Previous studies on most physiological signals including GSR have focused on either emotion detection or mental load measurement. This paper however studied the intersection of these two domains: it evaluated GSR features in detecting multiple cognitive load levels with affective interference.

Affective states are present in everyday life and their fluctuations have profound impacts on physiological data. As such, emotion interference is an important confounding factor which can vastly affect cognitive load measurement accuracy especially when using physiological data. This research, however, suggests that the studied electrodermal features can be robust against emotional changes and maintain their ability to objectively discriminate between different levels of cognitive demand while arousal is impacted by affective stimuli. Further investigation about the impact of affective stimuli on GSR features and simultaneous monitoring of other measures such as performance would expand our knowledge regarding robustness of electrodermal features for cognitive load measurement with affective interference.

Some limitations should be taken into account when interpreting the results of this study. A greater number of trials, especially in the first experiment, could enhance the certainty and power of findings. In the present research emotions were imposed by systematic and controlled use of visual stimuli; nevertheless visual interference and attentional regulation might have interfered with the measurement of cognitive load. The results may have improved by avoiding such potential confounding factors. In our future studies we will consider alternative methods of emotion induction such as presenting auditory stimuli to prevent visual interference and reduce/avert attentional regulation.

The explored features of the low-cost conveniently-captured GSR signal were found to detect multiple cognitive load levels significantly and with good classification accuracy in both studies even under emotional changes. Although some of the features performed better in either of the experiments, common promising ones also emerged. We are currently investigating feature fusion, combining features from similar and different behavioural and physiological modalities, to obtain higher performance in implicit cognitive load measurement. Furthermore, we aim to expand our research to explore cognitive load detection in more realistic application domains such as driving.

Monitoring mental state of the users can profoundly improve interactions and user experience in today's automated systems. Various practical aspects of people's life including, but not limited to, education, transportation (road, rail, sea or air), crisis and incident management, would dramatically benefit from objective, robust, accurate, real-time, unobtrusive measurement of cognitive load. In complex and critical tasks such as driving, piloting, air-traffic control and nuclear reactor management human's cognitive resources can easily be overloaded resulting in performance reduction, stress and mistakes with serious consequences. Thus, intelligent interface with a constant, personalized, situation-based adaptation can improve people's safety, interaction experience and performance. Findings of this research suggest the potential applicability of the proposed approaches in adaptable intelligent systems where unobtrusive monitoring of cognitive load would result in optimum interactions with humans.

ACKNOWLEDGMENTS

The authors would like to thank Mr Kun Yu and Mr Jianlong Zhou for discussions about statistical analysis.

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

REFERENCES

[1] J. L. Andreassi. 2007. Psychophysiology: human behavior and physiological response. Lawrence Erlbaum.

[2] C. Berka, D. Levendowski, M. Lumicao, A. Yau, G. Davis, V. Zivkovic, et al. 2007. EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks. *Aviation, Space, and Environmental Medicine, 78*(5), B231-44.

[3] W. Boucsein. 2012. *Electrodermal Activity*. Springer.

[4] Y. Cao, M. Theune, and A. Nijholt. 2009. Modality effects on cognitive load and performance in high-load information presentation. *International Conference on Intelligent User Interfaces*. New York, USA, 335-344.

[5] R. A. Calvo and D. Peters. 2014. Positive Computing: Technology for Wellbeing and Human Potential. MIT Press.

[6] P. Chandler and J. Sweller. 1996. Cognitive load while learning to use a computer program. *Applied cognitive psychology, 10*(2), 151–170.

[7] F. Chen, N. Ruiz, E. Choi, J. Epps, M. A. Khawaja, R. Taib, et al. 2012. Multimodal behaviour and interaction as indicators of cognitive load. *ACM Transactions on Interactive Intelligent Systems, 4*(2).

[8] S. Chen, J. Epps, N. Ruiz and F. Chen. 2011. Eye activity as a measure of human mental effort in HCI. *The 16th International Conference on Intelligent User Interfaces*. Palo Alto, California, USA: ACM Press, 315-318.

[9] N. Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87-114.

[10] D. E. Crundall, and G. Underwood. 1998. Effects of experience and processing demands on visual information acquisition in drivers. *Ergonomics, 41*(4), 448-459.

[11] M. E. Dawson, A. M. Schell and D. L. Filion. 2007. The electrodermal system. *Handbook of Psychophysiology*, 159-181.

[12] J. Engstrom, E. Johansson, and J. Ostlund. 2005. Effects of Visual and Cognitive Load in Real and Simulated Motorway Driving. *Transportation Research, 8*(2), 97-120.

[13] E. Haapalainen, S. Kim, J. F. Forlizzi, and A. K. Dey. 2010. Psycho-Physiological Measures for Assessing Cognitive Load. *International Conference on Ubiquitous Computing*. ACM Press, 26-29.

[14] S. Hart and L. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Human Mental Workload, 52*, 139-183.

[15] P. Hu, P. C. Ma and P. Y. Chau. 1999. Evaluation of user interface designs for information retrieval systems: a computer-based experiment. *Decision Support Systems, 27*(1), 125-143.

[16] M. S. Hussain, O. AlZoubi, R. A. Calvo and S. D'Mello. 2011. Affect Detection from Multichannel Physiology during Learning Sessions with AutoTutor. *International Conference on Artificial Intelligence in Education*. Auckland, New Zealand: Springer, 131-138.

[17] M. S. Hussain, R. Calvo and F. Chen. 2013. Automatic cognitive load detecting from face, physiology, task performance and fusion during affective interference. *Interacting with Computers, 25*(4).

[18] K. Huttunen, H. Keränen, E. Väyrynen, R. Pääkkönen, and T. Leino. 2011. Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights. *Applied ergonomics, 42*(2), 348-57.

[19] C. S. Ikehara and M. E. Crosby. 2005. Assessing Cognitive Load with Physiological Sensors. *Hawaii International Conference on System Sciences*. Hawaii, USA, 295a - 295a.

[20] L. C. Johnson and A. Lubin. 1966. Spontaneous electrodermal activity during waking and sleeping. *Psychophysiology, 3*, 8-17.

[21] P. Lang, M. Bradley and B. Cuthbert. 2008. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. University of Florida, Gainesville, FL.

[22] E. Leyman, G. Mirka, D. Kaber and C. Sommerich. 2004. Cervicobrachial muscle response to cognitive load in a dual-task scenario. *Ergonomics, 47*(6), 625-645.

[23] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche and B. Arnaldi. 2007. A review of classification algorithms for EEG-based brain-computer interfaces. *Journal of neural engineering, 4*(2), R1-R13.

[24] B. Mehler, B. Reimer, B. and J. F. Coughlin. 2012. Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: An on-road study across three age groups. *Human Factors: The Journal of Human Factors and Ergonomics Society, 54*(3), 396-412.

[25] K. R. Muller, M. Tangermann, G. Dornhege, M. Krauledat, G. Curio, and B. Blankertz. 2008. Machine learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring. *Journal of Neuroscience Methods, 167*(1), 82-90.

[26] S. Nakagawa. 2004. A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology, 15*(6), 1044-1045.
[27] A. Nakasone, H. Prendinger and M. Ishizuka. 2005. Emotion Recognition from Electromyography and Skin Conductance. *5th International Workshop on Biosignal Interpretation*. Tokyo, Japan, 219-222.
[28] N. Nourbakhsh, Y. Wang and F. Chen. 2013. GSR and blink features for cognitive load classification. *The 14th IFIP International Conference on Human-Computer Interaction.* Lecture Notes on Computer Science, Vol. 8117. Cape Town, South Africa: Springer, 159-166.
[29] N. Nourbakhsh, Y. Wang, F. Chen and R. A. Calvo. 2012. Using Galvanic Skin Response for Cognitive Load Measurement in Arithmetic and Reading Tasks. *Australian Computer-Human Interaction Conference*. Melbourne, Australia: ACM Press, 420-423.
[30] F. G. Paas. 1992. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*.
[31] F. Paas and J. Van Merriënboer. 1994a. Instructional Control of Cognitive Load in the Training of Complex Cognitive Tasks. *Educational Psychology Review, 6*(4), 351-371.
[32] F. Paas and J. Van Merriënboer. 1994b. Variability of Worked Examples and Transfer of Geometrical Problem-Solving Skills: A Cognitive-Load Approach. *Journal of Educational Psychology, 86*(1), 122-133.
[33] T. V. Perneger. 1998. What's wrong with Bonferroni adjustments. *British Medical Journal, 316(7139)*, 1236-1238.
[34] N. Ruiz, R. Taib, Y. Shi, Y. E. Choi and F. Chen. 2007. Using pen input features as indices of cognitive load. *The 9th International Conference on Multimodal Interfaces*. Nagoya, Aichi, Japan: ACM Press, 315-318.
[35] G. R. Saadé and A. C. Otrakji. 2007. First impressions last a lifetime: effect of interface type on disorientation and cognitive load. *Computers in Human Behavior, 23*(1), 525-535.
[36] W. Schnotz and C. Kürschner. 2007. A reconsideration of cognitive load theory. *Educational Psychology Review, 19*(4), 469–508.
[37] C. Setz, B. Arnrich, J. Schumm, R. La Marca and G. Tröster. 2010. Discriminating Stress from Cognitive Load Using a Wearable EDA Device. *Technology, 14*(2), 410-417.
[38] Y. Shi, N. Ruiz, R. Taib, E. H. Choi and F. Chen. 2007. Galvanic Skin Response (GSR) as an Index of Cognitive Load. *Computer Human Interaction 2007 Conference Work-in-Progress*. San Jose, California, USA: ACM Press, 2651-2656.
[39] R. M. Stern, W. J. Ray and K. S. Quigley. 2001. *Psychophysiological Recording.* Oxford University Press.
[40] J. Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.
[41] J. Sweller. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295–312.
[42] J. Sweller, J. J. Van Merrienboer and F. G. Paas. 1998. Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251-296.
[43] J. J. Vogel-Walcutt, J. B. Gebrim, C. Bowers, T. M. Carper and D. Nicholson. 2010. Cognitive load theory vs. constructivist approaches: which best leads to efficient, deep learning? *Journal of Computer Assisted Learning, 27*(2), 133-145.
[44] G. F. Wilson. 2011. An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures. *The International Journal of Aviation*, 37-41.
[45] J. Xu, Y. Wang, F. Chen, H. Choi, G. Li, S. Chen, et al. 2011. Pupillary Response Based Cognitive Workload Index under Luminance and Emotional Changes. *The ACM Conference on Human Factors in Computing Systems*. Vancouver BC, Canada: ACM Press, 1627-1632.
[46] B. Yin, F. Chen, N. Ruiz and E. Ambikairajah. 2008. Speech-based cognitive load monitoring system. *IEEE International Conference on Acoustics, Speech and Signal Processing*. Las Vegas, Nevada, USA, 2041-2044.
[47] K. Yu, J. Epps and F. Chen. 2013. Mental Workload Classification via Online Writing Features. *International Conference on Document Analysis & Recognition*. Washington, DC, USA, 1110 - 1114.

Received December 2015; revised July 2016; revised October 2016; revised February 2017; accepted March 2017