Radboud University Nijmegen

Bachelor Thesis

Artificial Intelligence

Classifying cognitive load from galvanic skin response time domain features

By:
Guido Zuidhof
guido.zuidhof@student.ru.nl
s4160703

Supervisors:
Dr. Louis Vuurpijl
Dr. Pashiera Barkhuysen
Dr. Ervin Poljac

January 26, 2015

Contents

Abstract

When performing a task, a certain amount of working memory is used. The amount of information a human is trying to process in working memory at any time is known as the cognitive load of a person. Biometrics can be used to determine this load. In this thesis galvanic skin response measurements were used to classify the cognitive load of a person. To study this, a pilot experiment was conducted where every participant performed four tasks with varying difficulty levels, each of which associated with a certain cognitive load level. An average 4-class classification accuracy of roughly 35%, and an average 2-class classification accuracy of around 62% was achieved.

Chapter 1

Introduction

When performing a task, a certain amount of working memory is used. The amount of information a human is trying to process in working memory at any time is known as the cognitive load of a person.

With the increasing capabilities of biometric measurement in smartwatches and smartphones it is exciting to see whether these biometric measures can be used to determine the cognitive load of a person. This could allow for dynamic adaptation of an application to best fit the mental state of the user.

This introduction starts with a brief overview of cognitive load theory, followed by an overview of various application areas where knowing the cognitive load of a person can be beneficial. Afterwards the methods of determining the cognitive load of a person is discussed, and the rationale of the chosen method studied in this thesis; a physiological measurement of the galvanic skin response.

1.1 Cognitive load

When processing information, we use working memory. The amount of memory used is not the same for every task. *Cognitive load theory* (Oviatt et al., 2004) revolves around the idea that (short-term) working memory is limited, whereas long-term memory is unlimited.

It is built upon *schema theory*, in which expertise in an area is a function of the acquisition of particular schemata (Bartlett and Bartlett, 1995). A schema is a mental structure to organise knowledge, (McVee et al., 2005) provides a more complete review of schema theory.

Cognitive load theory is nearly always studied in the context of learning new schemata. A schema is one element in this working memory, no matter how large it is (Mayer, 2014). The idea is that there are ways to reduce cognitive load by utilising long-term memory, and thus having more space in working memory, which would allow for more efficient learning.

Cognitive load theory distinguishes three types of cognitive load, which according to research can each be reduced in specific ways (Mayer, 2002):

Extraneous cognitive load This cognitive load is influenced by how the material or task is presented. It can be reduced by using worked-out examples, using diagrams, or presenting the material in multiple modalities (Mousavi et al., 1995).

Intrinsic cognitive load This cognitive load is determined by the complexity of the material at hand. It can be reduced by splitting up the task or building upon former (informal) knowledge.

Germane cognitive load Germane cognitive load is the cognitive load associated with the process of building, acquiring and automating new schemas. The greater the space in working memory for this load (by reducing the intrinsic and extraneous cognitive load), the greater the ease of learning.

1.2 Applications

There are many application areas where knowing the cognitive load of a person yields valuable information. Here I will highlight three:

Learning Most applications and research regarding cognitive load theory is applied in the context of multimedia learning (Brunken et al., 2003). The goal in this domain is to find the way to structure and present information to ease learning the most. In other words, to reduce the intrinsic and extraneous cognitive load as much as possible, so that more working memory is left for germane cognitive load. Generally speaking, presenting information in more modalities in multimedia learning tasks decreases the intrinsic and extraneous cognitive load of the task. An example is the use of figures and graphs to visualize information explained in text.

Usability testing Cognitive load measures, and more generally stress and arousal measures, are used by human computer interaction (HCI) developers and researchers to evaluate the usability of software or systems (Jacob and Karn, 2003) (Schmutz et al., 2009).

Schmutz et al. (2009) investigated cognitive load measurement in web applications. Concluded was that cognitive load measures can be used as a *"valuable additional measure of efficiency"*, which strongly correlated to general satisfaction of the user.

Mental workload ¡TODO uitleg wat mental workload vs cognitive load¿

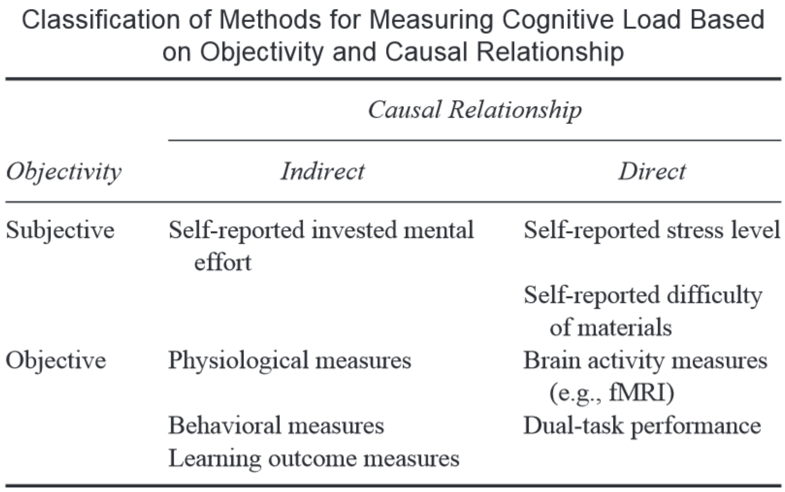
The interaction of mental workload and the three types of cognitive load has been studied by Galy et al. (2012). Increasing the cognitive load, for instance by increasing the difficulty of the task, was shown to have an additive interaction with mental workload.

Knowing the cognitive load of a person can be used as an indicator for the mental workload of the person. Dynamically using this information to scale up or down the cognitive load of an application can be used to prevent mental overload, which can lead to accidents at work.

1.3 Determining cognitive load

There are many approaches to determining cognitive load. The figure below, from Brunken et al. (2003), puts the different methods in a matrix based on their objectivity and causal relationship between measurements and cognitive load. Every end of this spectrum has it's benefits and downsides, for instance, subjective methods are a good way to measure the cognitive associated with any task by having the subject fill out a questionnaire. They are however self-reported measured, and thus subjective, which means that the honesty and quality of introspection of subjects may play a role in the results.

On the other end, there are the objective measures, these generally require a more careful design of experiments. The benefit is that they are objective, which means it is not subject to the issues associated with subjective measures described above. In this thesis the indirect, objective method of *physiological measures* is studied.



Classification of Methods for Measuring Cognitive Load Based on Objectivity and Causal Relationship

	Causal Relationship	
Objectivity	*Indirect*	*Direct*
Subjective	Self-reported invested mental effort	Self-reported stress level
		Self-reported difficulty of materials
Objective	Physiological measures	Brain activity measures (e.g., fMRI)
	Behavioral measures	Dual-task performance
	Learning outcome measures	

Figure 1.1: Methods of determining cognitive load. (Brunken et al., 2003).

One of the main benefits of this approach is the objectivity and the possibility of using the cognitive load measurement dynamically, whereas with other objective measures the complete task has to be completed and the results of which analysed. Downsides of this approach are the possibility of noisy data, sensors may have slight defects and what one measures may be only partially influenced by the cognitive load of a person. Also, the participants have to wear sensors or one has to be in front of a sensor (such as a camera for eye blink rate measurement), which may impair movement.

Psychiological measures such as heart rate, blink rate, pupil dilation, skin conductance and more have been shown to have a correlation to cognitive load. In this thesis skin conductance was the biometric of choice. In the next section I will explain the rationale behind the decision.

1.3.1 Galvanic Skin Response

Galvanic Skin Response (GSR) is an interesting physiological response to measure as sensors are cheap and simple. Also, the subject is not or marginally restricted in movement, this allows for long-term measurement (all day), if need be outside of the lab. To this end GSR is the biometric of choice in this explorative study.

Because of these features mobile GSR measurement systems, for instance around the foot (Gravenhorst et al., 2013) are being developed to measure (patients) throughout the day. Also, new devices such as smartwatches are fitted with GSR electrodes increasingly often, which means findings could be applied there. An example would be using GSR data measured by a smart watch to dynamically change an application's presentation of information.

GSR can be measured in two ways, by measuring conductance (in siemens) or resistance (in ohms). Measuring skin conductance *"bears a simpler more linear relationship to the underlying processes of psychological interest than its repriprocal, SR [skin resistance]"* (Lykken and Venables, 1971), therefore any GSR measurement in this thesis is done by measuring skin conductance.

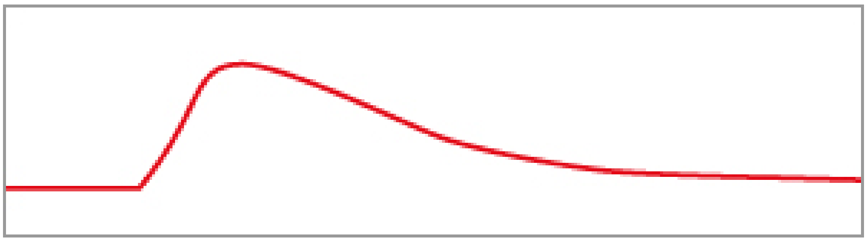


Figure 1.2: Typical skin conductance response (SCR) (Dow, 2015).

Above is a figure showing the typical skin conductance response to a discrete stimulus. Note the response onset, this onset can be up to 4 seconds from the stimulus onset (Dow, 2015). After the peak, an exponential decay is observed.

1.4 State of the art

Galvanic skin response readings have been studied in regard to cognitive load.

Shi et al. (2007) investigated the possibility of using GSR as an index of cognitive load during the use of unimodal and multimodal versions of the same interface. It has been shown that multimodal interfaces *"support the user in managing cognitive load"* (Oviatt et al., 2004), and thus multimodal interfaces have a lower corresponding cognitive load.

Shi et al. found that GSR values increase when cognitive load levels increase across all subjects, and conclude that GSR can be used to "serve as an objective indicator of [a] user's cognitive load".

Nourbakhsh et al. (2012) showed that both time and frequency domain features from GSR data significantly correlate with the cognitive load of a person.

In the follow-up research GSR and blink features were used for cognitive load classification (Nourbakhsh et al., 2013). Combining these two featuresets, around 75% accuracy for binary classification, and 50% accuracy for four-class classification was achieved. Only GSR features lead to accuracies of around 66% and 35% for binary and 4-class classification resepctively.

Xu (2014) analyzed the mental stress and workload using both galvanic skin response and heart rate variability. The aim was to find the correlation between mental stress and design activities. These activities included designing various objects, such as a house that can easily fly to another place. The results showed that the mean GSR levels were significantly higher during the task compared to the task, and the post-test resting state GSR was also significantly higher than the pre-test resting state level.

Novel ways of measuring skin conductance are also studied. Kim et al. (2014) used a conductive, flexible polymer foam material to create a sensor that measures the signals from the back of the user. This sensor was easily attached and detached from ordinary clothes, and a correlation to more ordinary GSR sensors (finger electrodes) proved to be high enough to show the feasibility of this new way of measuring GSR.

1.5 Research aim

The goal of this thesis is to investigate the possibility of determining what the cognitive load of a person is, given GSR measurements. An explorative experiment will be used with different tasks, each of which evokes a different level of cognitive load. From this data a classifier is built, which can classify which task the user was doing based on features derived from GSR, and thus say something about the cognitive load of a person.

The research question reads:

Can cognitive load be classified from galvanic skin response measurements?

1.5.1 Organization of this thesis

In the next chapter I will explain the method with which this research question is answered. Then, in chapter three, the results are presented, of which the implications are discussed in the final chapter.

Chapter 2

Method

2.1 Experiment

The experiment is mostly based on the arithmetic task experiment conducted by Nourbakhsh et al. (Nourbakhsh et al., 2012). The experiment involves adding up four numbers, displayed one by one, and selecting an answer from three possible answers.

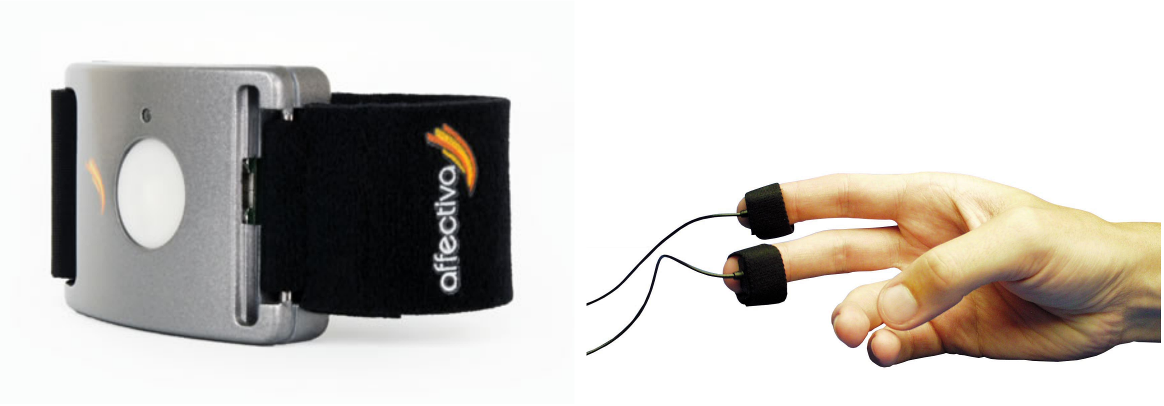
There are four difficulty levels of this task. Difficulty level one involves adding up four binary numbers (1 and 0). Difficulty level two, three and four consist of adding up four numbers with a length of one, two and three digits respectively.

The subject is first shown an amount of stars corresponding to the difficulty level for 8 seconds, after which the numbers are shown one by one for 4 seconds. Three answers are then presented on the screen, of which one is correct. The subject subsequently selects the correct answer using a standard computer mouse. There was no time limit for selecting an answer.

In (Nourbakhsh et al., 2013) these tasks were shown to have significant difference in subjective ratings of task difficulty levels, as well as a siginificant difference in response time of different task levels. From which is shown that these designed tasks manipulate cognitive load effectively.

2.1.1 Apparatus

The task was displayed in a 15.6" laptop screen, input was given using a standard computer mouse. Two galvanic skin response sensors were used in this experiment. The participants wore both sensors at the same time on the left arm. Below, the two sensors are briefly discussed.



(a) Affectiva Q Sensor

(b) BIOPAC MP30 finger electrodes

Figure 2.1: Sensors used in the experiment

Affectiva Q Sensor This sensor is embedded in a wrist band, it uses a dry electrode. The sensor side of the wrist was placed on the bottom of the wrist. The sampling rate of this sensor was set to 32 samples per second. This is the highest possible sampling rate setting of this device.

BIOPAC MP30 The BIOPAC MP30 was used in conjunction with finger GSR sensors. This sensor wraps around two fingers, in this experiment the index and ring finger. It uses gel electrodes, Grass EC33 Electrode Paste was used for this purpose.

The sampling rate was set to 1000 samples per second. This is an arbitrarily chosen high sampling rate, much higher than the frequencies at which Nourbakhsh et al. (2012) found significant differences in the power spectrum between the different tasks.

2.1.2 Participants

Six participants took part in this experiment, five male and one female. Their age ranged between 19 and 25 at the time of the experiment. They were not compensated for their participation.

2.1.3 Experimental Design

The task consisted of 8 arithmetic tasks, with 4 difficulty levels. Two tasks of every difficulty level were completed by the participants, in a random order.

A single task consisted of adding up four numbers, and selecting an answer answer from three options. The numbers to add were shown one by one, for four seconds. Before a task starts, a number of stars is shown equal to the difficulty level, for eight seconds. The tasks followed each other without breaks. There was no time limit for selecting the answer, there was no feedback whether the selected answer was correct.

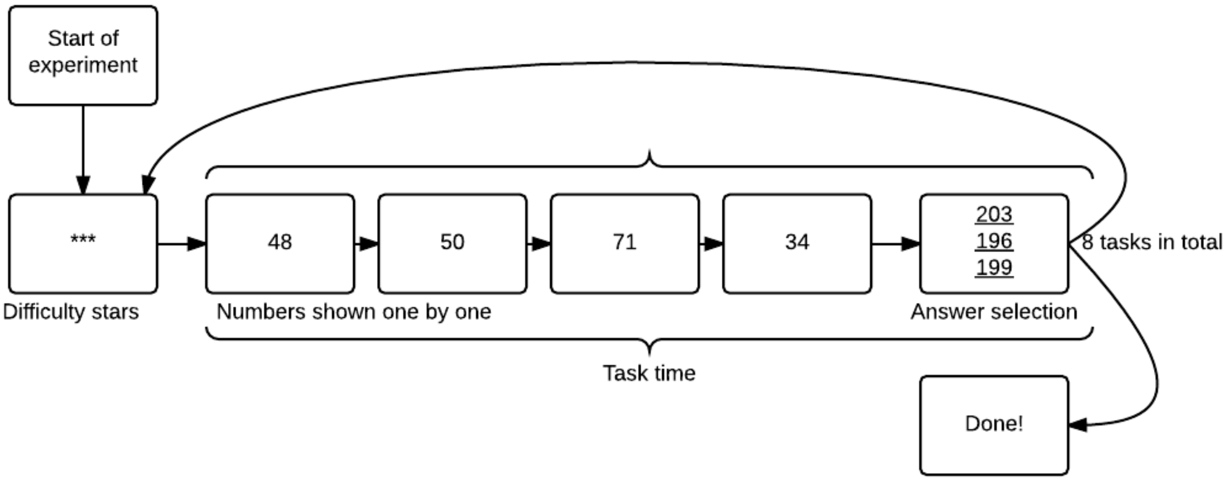


Figure 2.2: Experiment structure

2.1.4 Procedure

The participants were first explained the task and seated in front of the laptop. Then the sensors were applied to the left hand and wrist. The participant was then instructed to place the left hand on the table in front of him/her and not to move it, in order to prevent possible motion artifacts which have been shown to influence GSR readings (Chellali and Hennig, 2013).

2.2 Analysis

2.2.1 Affectiva Q Sensor Data

The data gathered from the Affectiva Q Sensor proved to be unusable. In most participants no proper skin contact seemed to have been made during the full experiment. Dry GSR sensors need some sweat to ensure a good connection with both electrodes. Given the short experiment duration (a few minutes) this connection was possibly never established.

In hindsight the participants could have worn the device longer before the first task started, or they could have been asked to perform some light exercise.

Therefore, any data discussed from here on is solely that which has been collected using the BIOPAC MP30 and it's finger electrodes.

2.2.2 Preprocessing

All data files were preprocessed using custom scripts written in *MATLAB (The MathWorks, Inc.)*. The collected data contained a DC current, which was smoothed by filtering and resampling to 100 samples per second using a Savitzky-Golay filter (Savitzky and Golay, 1964).

The data per subject was then cut into separate frames, the tasks. A task (slice of data) runs from the time that the first number is shown, to the time that an answer is given. These tasks were then smoothed one last time, to get rid of a small remaining amount of DC noise. This was again done using a Savitzky-Golay filter.

The data was then normalized per subject over all datapoints in all tasks. This was done by dividing every data point of the subject by the average of the subject over all tasks.

Let the average over all tasks:

$$AverageGSR(s) = \frac{\sum_{q \in Q_s} \sum_{t \in T_q} GSR(s, q, t)}{\sum_{q \in Q_s} |T_q|}$$

With s a given subject, Q_s a set of all tasks performed by s, and T_q the set of all measurements in task q, and $GSR(s, q, t)$ be the GSR datapoint of subject s, task q at time t.

The normalised GSR is then given by:

$$NormGSR(s, q, t) = \frac{GSR(s, q, t)}{AverageGSR(s)}$$

Note that the normalization is done by dividing by the average of the subject over all tasks, and not by simply transforming all measurements to be between 0 and 1. What (original) GSR value would become 1 and 0 would then be quite arbitrary per subject, and not all subjects would be on the same scale, as outliers can totally influence this range.

This normalised GSR was then cut into the specific tasks.

2.2.3 Transformation

From the measurements from every task certain time domain features were extracted.

The extracted features:

- Average over all points in the task.
- Accumulative of all points in the task.
- Standard deviation of all points in the task.
- Difference between first and last point of the task.
- Peaks with varying tresholds of what a peak is.

Below I will elaborate how these features were calculated from the data. In these formula's s is a subject, q is a task and T_q are all the measurements (datapoints) associated with task q.

Average

$$AvgTaskGSR(s,q) = \frac{\sum_{t \in T_q} NormGSR(s,q,t)}{|T_q|}$$

Accumulative

$$AccTaskGSR(s,q) = \sum_{t \in T_q} NormGSR(s,q,t)$$

Standard deviation

$$StdDevTaskGSR(s,q) = \sqrt{\frac{1}{|T_q|} \sum_{t \in T_q} (NormGSR(s,q,t) - AvgTaskGSR(s,q))^2}$$

Difference

$$DiffTaskGSR(s,q) = NormGSR(s,q,t_{|T_q|}) - NormGSR(s,q,t_1)$$

where t_1 is the first measurement in T_q, and $t_{|T_q|}$ the last measurement of the task.

Peaks For peak detection the *PeakFinder* script (Yoder, 2009), found in the *Matlab Central File Exchange* was used under BSD License. This script finds local maxima or minima, even in noisy signals, and can be supplied with a treshold that determines what classifies as a peak.

This script was called with four different threshold levels, given by

$$Threshold(s,q,k) = \frac{\max_{t \in T_q} NormGSR(s,q,t) - \min_{t \in T_q} NormGSR(s,q,t)}{(2k)^2}$$

with $k \in \{1,2,3,4\}$.

2.2.4 Data mining

These extracted features are then fed into *Weka 3* (Hall et al., 2009), which is data mining software in Java.

Three of the most commonly used classification algorithms were used (Ruparel et al., 2013). Namely a Bayesian classification algorithm (*NaïveBayes*, which is an implementation based on (John and Langley, 1995)), a support vector machine (SVM) classification algorithm (*LibSVM*) (Chang and Lin, 2011), with SVM model type C-SVC and a radial basis kernel type, and a neural network classifier (*MultilayerPerceptron*, with 8 hidden layers).

Two classifiers were trained with each algorithm, one for 4-class classification, and one for 2-class (binary) classification (with task difficulty 1 and 2 combined, and 3 and 4 combined).

Chapter 3

Results

3.1 Gathered data

Below in figure 3.1 the raw data is shown from one of the subject over the course of the whole experiment (including some trailing seconds of disconnecting the measurement device). In figure 3.2 this data has been filtered using the first Savitzky-Golay filter.

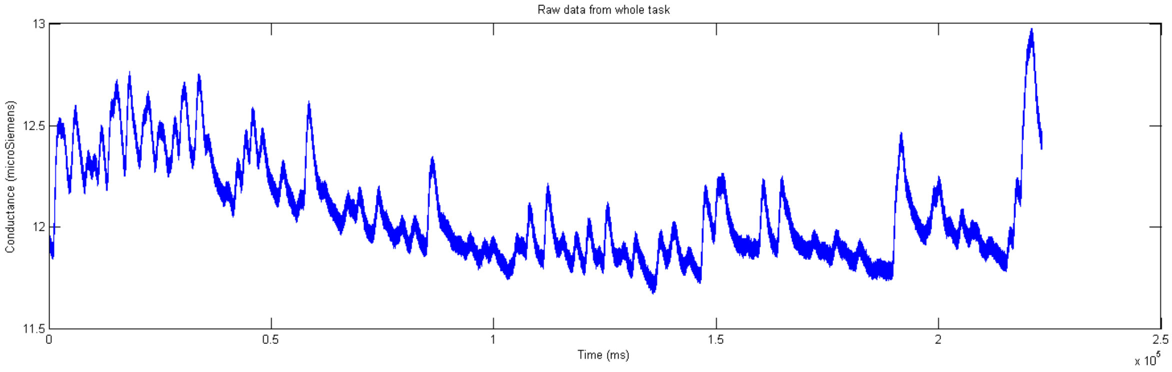


Figure 3.1: Raw data from a single participant.

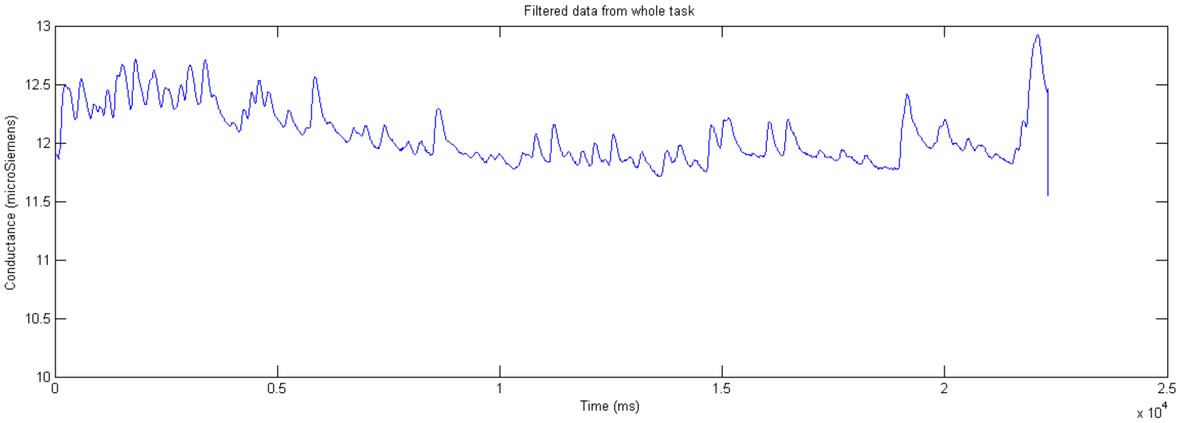
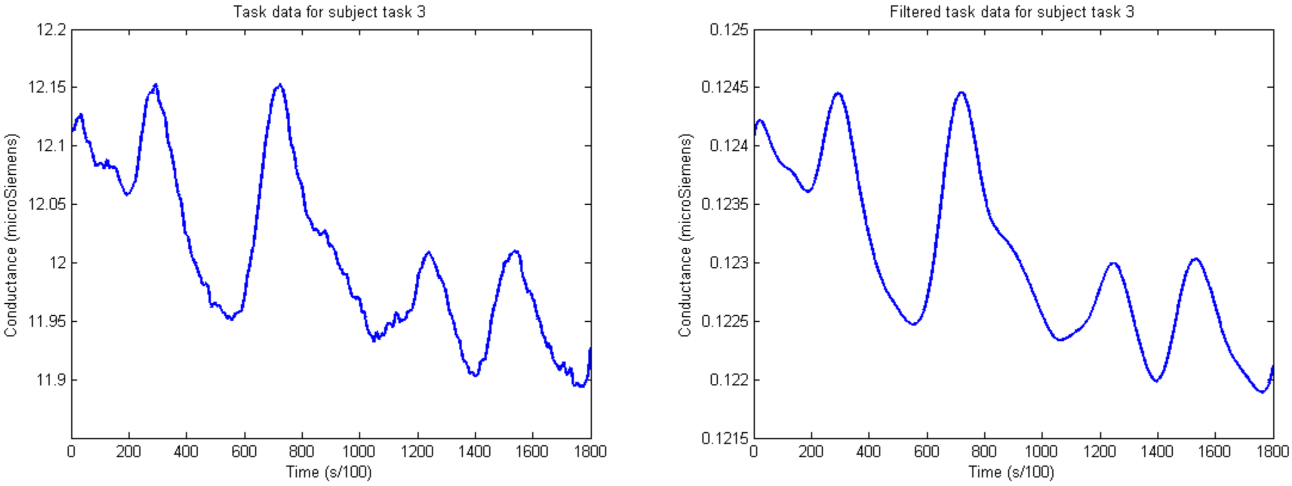


Figure 3.2: Data from a single participant after Savitzky-Golay filter.

A typical task is shown in figure 3.3a, this is a slice of the data in plotted in figure 3.2. In 3.3b this data has been normalized as described in the previous chapter, as well as filtered using a Savitzky-Golay filter to smooth over any remaining noise.



(a) unnormalized

(b) filtered and normalized

Figure 3.3: Task 3 (difficulty 2) of subject 2

In the appendix of this thesis there are some visualizations of characteristics of the processed gathered data.

3.2 Classification accuracy

Table 3.1: Classification accuracy

Classification Algorithm	2-Class Classification	4-Class Classification
NaiveBayes	62.5%	35.4%
LibSVM	68.8%	22.9%
MultilayerPerceptron	56.3%	41.7%

The performance of each predictive model was estimated using 10 fold cross-validation. The percentages in the table are the percentage of correctly classified task difficulties by the predictive model.

3.3 Confusion matrices

Table 3.2: NaiveBayes confusion matrix

1	2	3	4	<- Classified as difficulty
1	7	1	3	Difficulty 1
3	7	0	2	Difficulty 2
2	4	2	4	Difficulty 3
2	2	1	7	Difficulty 4

Table 3.3: LibSVM confusion matrix

1	2	3	4	<- Classified as difficulty
1	5	6	0	Difficulty 1
5	3	3	1	Difficulty 2
3	1	5	3	Difficulty 3
5	0	5	2	Difficulty 4

Table 3.4: MultilayerPerceptron confusion matrix

1	2	3	4	<- Classified as difficulty
3	4	3	2	Difficulty 1
3	5	2	2	Difficulty 2
2	3	6	1	Difficulty 3
2	1	3	6	Difficulty 4

These confusion matrices visualize the performance of the classification algorithm. From a confusion matrix one can see how often a class (difficulty of task) is predicted as another class. The correctly classified instances are found on the diagonal (top-left to bottom-right).

Chapter 4

Discussion

To rehash, the research question reads *Can cognitive load be classified from galvanic skin response measurements?*. In this section I will attempt to answer this question with the results from this experiment.

4.1 Interpretation

The results show that the classification accuracy of the classifiers is better than chance dictates for two out of three used classification algorithms, however not much. In the binary classification task the results were more reasonable, where every classifier performs better than chance. These classification rates are slightly lower than those reported by Nourbakhsh et al. (2013) in a similar experiment, this may be due to the smaller dataset.

When considering the confusion matrices of the 4-class classification it can be noted that the task most often classified as having a wrong difficulty level was task 1, which was more often classified as difficulty 2 and 3 than 1. In fact, it has only been classified correct once out of twelve entries in the NaiveBayes and LibSVM model, and a mere three times in the MultilayerPerceptron model. Perhaps this task was too easy and caused participants to become distracted.

There are many possible explanations why this classification accuracy is not greater than it is. The first explanation is that from just GSR data, it simply can not be predicted any better. Next I will highlight some other possible reasons:

Loss of information from filter

Filtering the data, to remove the DC current noise, lead to some data loss. A Savitzky-Golay filter smooths peaks slightly, which may have influenced the classification rate.

Task bleeding

Another explanation may be the bleeding of tasks into other tasks. If for instance a subject first gets two hard tasks, and then an easy one, it is imaginable that the subjects GSR value for that easy task is higher than it would have been without first having done harder tasks. When a stimulus is presented to a subject, there is an exponential decay in GSR value. There may have not been enough time between the tasks for the GSR value of the person to return to a neutral state.

This bleeding effect is mitigated by the random order in which the tasks are presented and the time between tasks, however with the relatively small subject pool it is not unimaginable this may have had its effect on the classification accuracy.

Dataset size

With 6 subjects and 8 tasks per subject, the total dataset consisted of 48 entries. Training a classification algorithm on a small dataset may cause outliers to influence it to fairly great effect.

Task cognitive load manipulation

Despite Nourbakhsh et al. (2013) showing that the tasks effectively manipulate the cognitive load of the subject, this was concluded from subjective reports and response times. It may be that these response times were manipulated by something other than cognitive load, and that the tasks do not evoke a difference in cognitive load in the subject. This could be the case between all difficulties, or between some. It may also be that this difference in cognitive load is very small, making classification harder, especially with more than two classes.

Time domain features

All extracted features were solely features from the time domain, however frequency domain features were shown by (Nourbakhsh et al., 2012) to significantly correlate with the task difficulty as well. Adding frequency domain features to the featureset, may improve classification rates.

4.2 Future research

Future research could be measuring GSR data using longer, carefully designed tasks, of which the cognitive load can be determined using an established cognitive load index. Also future research could include investigations of how classification accuracy changes when using data measured by more portable (consumer) sensors, whose measurements are likely of lesser quality than the used lab equipment in this experiment.

There are benefits to using these consumer sensors. Generally, these sensors are much cheaper, allowing for more simultaneous participants (which is especially useful in long-term measurements, for instance over the course of a week). This can promote the ecological validity of the test, as these sensors can be worn during real world activities. Also, participants may already own these devices, and participating could be as simple as downloading an application from the *App Store*.

Appendix A

Visualization of extracted features

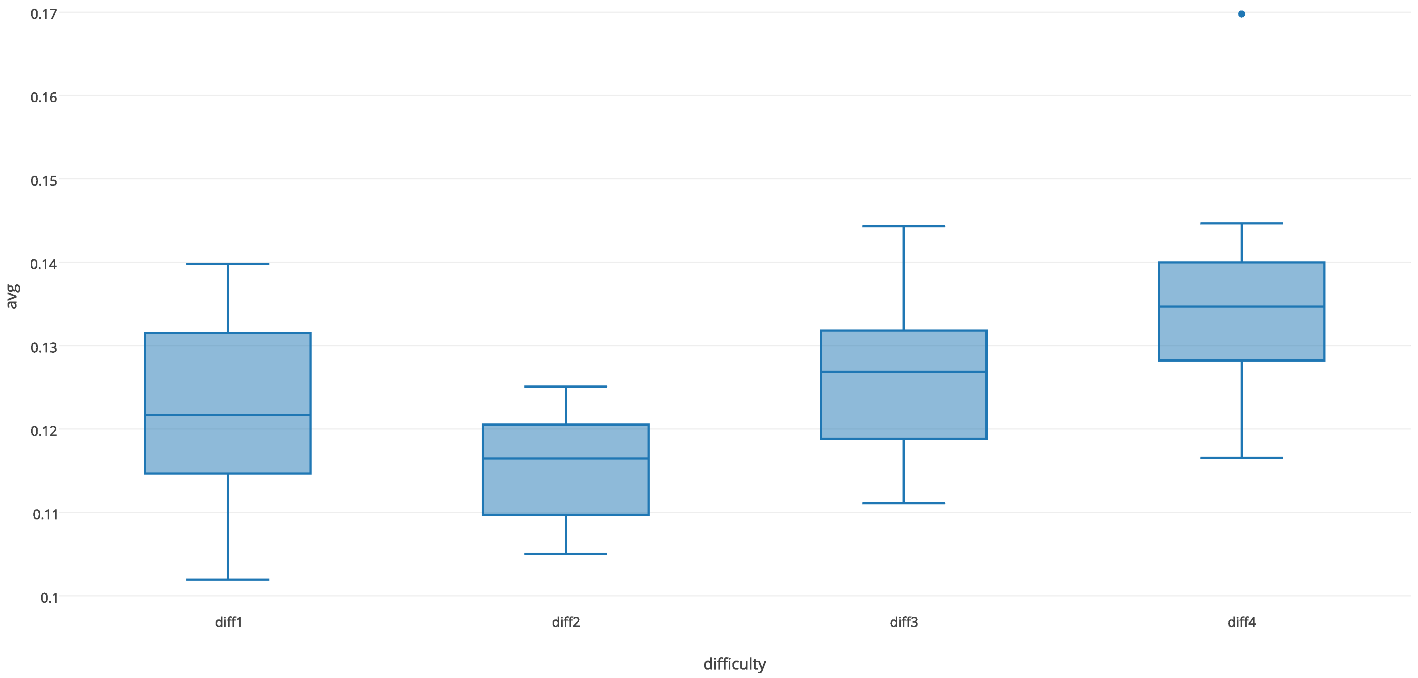


Figure A.1: Average GSR.

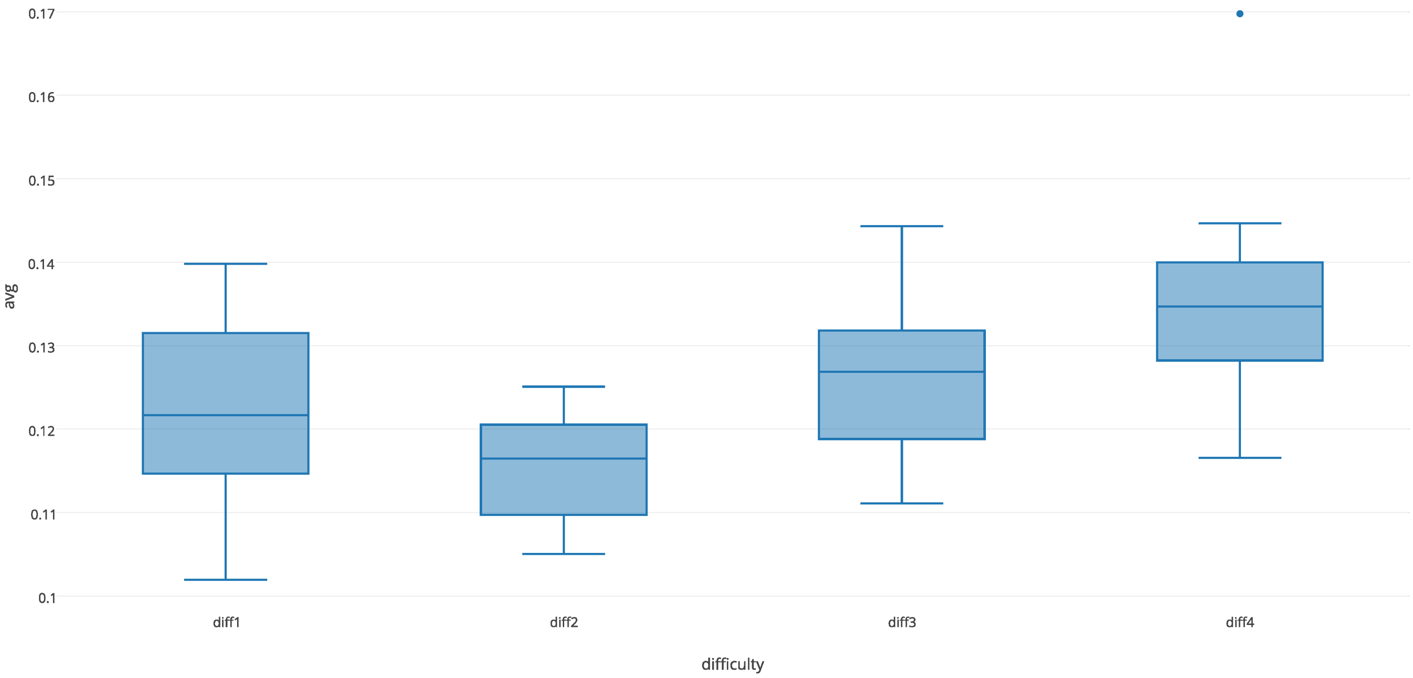


Figure A.2: Standard deviation.

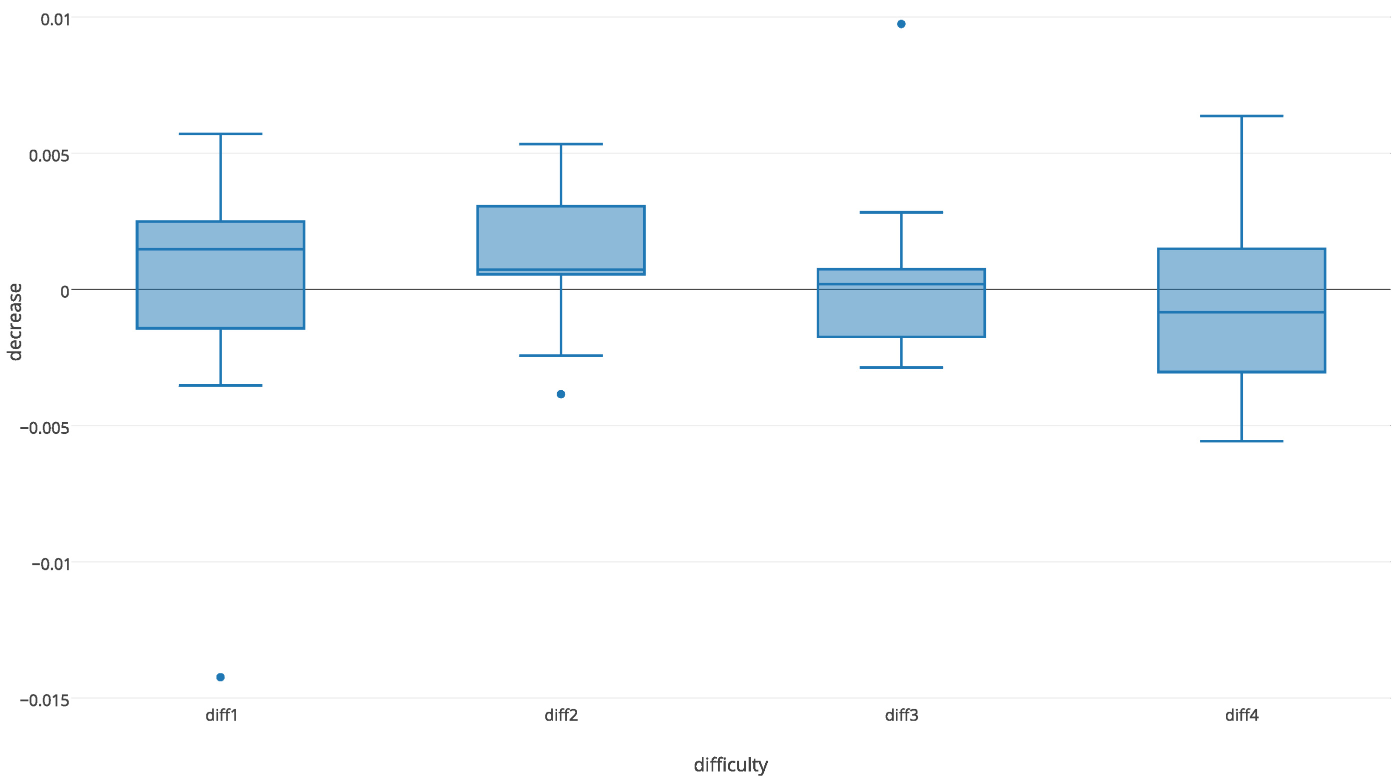


Figure A.3: Decrease in GSR over task.

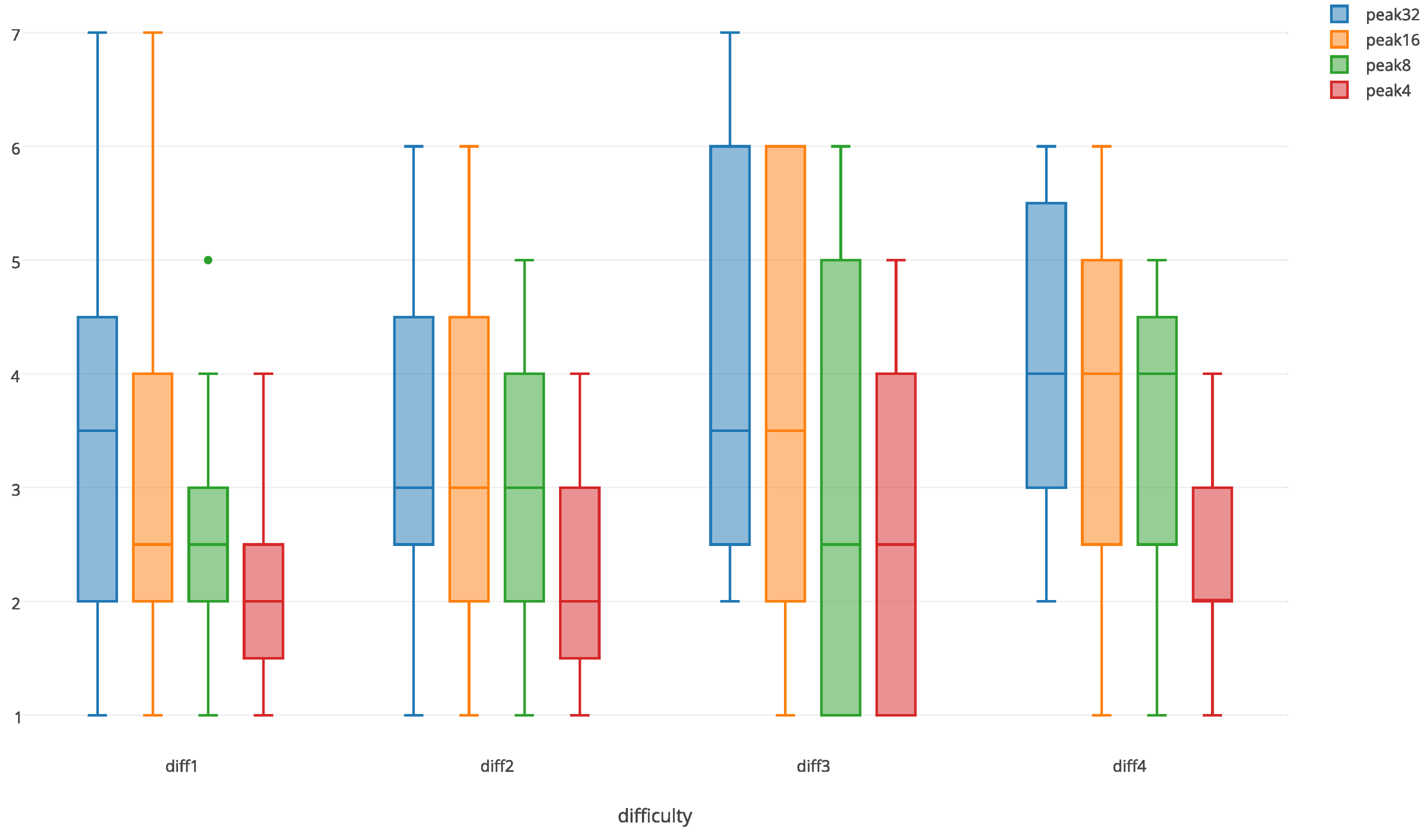


Figure A.4: Peak count with various tresholds.

Bibliography

Bartlett, F. C. and Bartlett, F. C. (1995). *Remembering: A study in experimental and social psychology*, volume 14. Cambridge University Press.

Brunken, R., Plass, J. L., and Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, 38(1):53–61.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27.

Chellali, R. and Hennig, S. (2013). Is it time to rethink motion artifacts? temporal relationships between electrodermal activity and body movements in real-life conditions. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 330–335.

Dow, R. (2015). SC_explained.

Galy, E., Cariou, M., and Mlan, C. (2012). What is the relationship between mental workload factors and cognitive load types? *International Journal of Psychophysiology*, 83(3):269 – 275.

Gravenhorst, F., Muaremi, A., Tröster, G., Arnrich, B., and Gruenerbl, A. (2013). Towards a mobile galvanic skin response measurement system for mentally disordered patients. In *Proceedings of the 8th International Conference on Body Area Networks*, BodyNets '13, pages 432–435, ICST, Brussels, Belgium, Belgium. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Jacob, R. J. and Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind*, 2(3):4.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc.

Kim, J., Kwon, S., Seo, S., and Park, K. (2014). Highly wearable galvanic skin response sensor using flexible and conductive polymer foam. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 6631–6634.

Lykken, D. T. and Venables, P. H. (1971). Direct measurement of skin conductance: A proposal for standardization. *Psychophysiology*, 8(5):656–672.

Mayer, R. (2014). *The Cambridge Handbook of Multimedia Learning.* Cambridge Handbooks in Psychology. Cambridge University Press.

Mayer, R. E. (2002). Multimedia learning. *Psychology of Learning and Motivation*, 41:85–139.

McVee, M. B., Dunsmore, K., and Gavelek, J. R. (2005). Schema theory revisited. *Review of educational research*, 75(4):531–566.

Mousavi, S. Y., Low, R., and Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of educational psychology*, 87(2):319.

Nourbakhsh, N., Wang, Y., and Chen, F. (2013). Gsr and blink features for cognitive load classification. In *INTERACT 2013*, page 8, Cape Town, South Africa.

Nourbakhsh, N., Wang, Y., Chen, F., and Calvo, R. A. (2012). Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th Australian Computer-Human Interaction Conference*, OzCHI '12, pages 420–423, New York, NY, USA. ACM.

Oviatt, S., Coulston, R., and Lunsford, R. (2004). When do we interact multimodally?: Cognitive load and multimodal communication patterns. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, pages 129–136, New York, NY, USA. ACM.

Ruparel, N. H., Shahane, N. M., and Bhamare, D. P. (2013). Article: Learning from small data set to build classification model: A survey. *IJCA Proceedings on International Conference on Recent Trends in Engineering and Technology 2013*, ICRTET(4):23–26. Full text available.

Savitzky, A. and Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639.

Schmutz, P., Heinz, S., Métrailler, Y., and Opwis, K. (2009). Cognitive load in ecommerce applications: Measurement and effects on user satisfaction. *Adv. in Hum.-Comp. Int.*, 2009:3:1–3:9.

Shi, Y., Ruiz, N., Taib, R., Choi, E., and Chen, F. (2007). Galvanic skin response (gsr) as an index of cognitive load. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '07, pages 2651–2656, New York, NY, USA. ACM.

Xu, X. (2014). Analysis on mental stress?workload using heart rate variability and galvanic skin response during design process.

Yoder, N. (2009). Peakfinder. Downloaded from the Matlab Central File Exchange.