Predicting a Pilot's Cognitive State from Physiological Measurements

J.A. Crijnen
STUDENT NUMBER: 2031888
ADMINISTRATION NUMBER: 666280

THESIS SUBMITTED IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:

dr. H. J. Brighton

Second Reader: dr. J. S. Olier Jauregui

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
June 2019

Preface

I thank dr. H. J. Brighton for his guidance, advice and patience during this research.

J.A. Crijnen

June 2019

Abstract

This thesis investigates the cognitive states of pilots during experiments, using data which relate to physiological attributes. Although expired, a challenge was initiated by consulting firm Booz Allen Hamilton, during which data scientists were encouraged to build a model with detection capabilities to prevent aviation accidents and incidents. The overarching question this thesis investigates is: Can the cognitive state of a pilot be predicted from physiological measurements?
Previous research has found that cognitive state detection can improve aviation safety. During this research the Human Factor Classification and Analysis System was consulted to place the findings in a broader perspective. The task investigated in this thesis is two-folded: (1) cognitive state classification and (2) cognitive state change detection. The key problem which is addressed, is the engineering of features from complex data. Therefore frequency domain analysis and sliding window time analysis are performed. Out of 73 variables, the five variables that contribute the most to the model's performance are selected. The proposed model achieves an F1 score of 0.67 in detecting the appropriate cognitive state for a specific pilot in the test data. The average F1 score on the testing data is 0.55, which is higher than the benchmark model without engineered features (0.48). Especially during startle surprise and diverted attention classification, lower performance is observed. In addition, not all pilots show predictive characteristics. The model reveals potential success in the field of cognitive state prediction, and therewith increasing aviation safety. For further application, improvement in performance is necessary. Cognitive state prediction appears to be complex, nevertheless, it could hold the key to further increase the safety within aviation.

Contents

Predicting a Pilot's Cognitive State from Physiological Measurements

J.A. Crijnen

1. Introduction

Although operations within aviation become increasingly automated, a substantial part of the decision making in the aircraft's cockpit is human based (Kelly and Efthymiou 2019). Decades ago, the possibilities of Artificial Intelligence (AI) in aviation were discovered. For example, automating Air Traffic Control (ATC) (Gosling 1987). Gosling refers to Rich's definition of Artificial Intelligence as being "...the study of how to make computers do things at which, at the moment, people are better", and he expected a wide range of possibilities for this development in aviation (Rich 1983). In 1994, the U.S. Department of Transportation published a document indicating various aviation based possibilities of AI (Harrison, Saunders, and Janowitz 1994). In addition to ATC, maintenance, air space efficiency, reducing flight costs and pilot decision making were mentioned promising applications. The latter is what this thesis investigates. Because of the increasing amount of knowledge in the field of AI, this research aims to increase the safety of aviation by means of improving the decision making process. This can be achieved by informing the flight crew of their cognitive state. Oster, Strong, and Zorn (2013) argue for switching to a proactive strategy with regards to aviation safety:

> The next generation of safety challenges now requires development and understanding of new forms of data to improve safety in other segments of commercial aviation, and moving from a reactive, incident-based approach toward a more proactive, predictive and systems-based approach. (p. 163)

The necessity to improve aviation safety does not need further explanation.

1.1 Research

This thesis investigates the cognitive states of pilots using physiological data provided by Kaggle (Kaggle 2018). Although expired, a challenge was initiated by consulting firm Booz Allen Hamilton, where data scientists were encouraged to build a model with detection capabilities, to prevent accidents and incidents (Booz Allen Hamilton 2019). At the time of writing this thesis, no winner has been appointed yet.
The subtitle of the data information web-page by Kaggle states: "Can you tell when a pilot is heading for trouble?". This seems to define the problem statement concisely, although perhaps, it might be too optimistic and informally formulated. A more scientific approach is to investigate if the cognitive state of a pilot can be predicted from physiological measurements.

1.2 Problem statement

The research investigates the following overarching questions.

RQ: *Can the cognitive state of a pilot be predicted from physiological measurements?*

This is broken down into the following sub-questions.

SQ1: *What is the relative contribution of the physiological features when predicting a pilot's cognitive state?*

SQ2: *Are some cognitive states easier to predict than others?*

SQ3: *Are all pilots suitable for the application of cognitive state prediction?*

These sub-questions consider the predictive capabilities of the features. Because of the complex dataset, the research focuses on which features can be engineered to enable the training of predictive models. SQ1 discusses the contribution of the dependent variables to the predictions on the test data. Additionally, this contains the level of difficulty in applying cognitive state prediction in flight operations. If the situation arises that relatively basic physiological measurements contain decent predicting capabilities, then practical issues regarding the montage of equipment to pilots, will be less of a limitation. For example, wearing heart rate detection sensors has relatively low impact compared to equipment for brain activity registration.
SQ2 explains the performance of the proposed model to recognize each cognitive state. Presumably, not all cognitive states occur as frequently as others. Because no distinction is made in importance of recognizing a certain cognitive state, identifying the less frequent cognitive states might be challenging. This thesis stresses to prevent the model to be only accurate. It should also be unbiased in the prediction of less frequent cognitive states. This performance trade-off is clarified in more detail in section 3.5.
To entirely place the results in perspective it is important to investigate the generalisability of the model, which is specified in SQ3. To enable cognitive state prediction within a broad range of human physiological characteristics, overfitting of a model trained on few pilots is a high risk (Obermeyer and Emanuel 2016). Additionally, complex models tend to train on noise (besides the desirable signal), which is expected to be present in the consulted dataset (Skocik et al. 2016).
Finally, the results of the experiments will be discussed to place the findings in a broader perspective with regards to aviation safety. To understand in detail the impact of possible cognitive state detection, it is investigated in which manner aviation safety can benefit from this application. The Human Factor Analysis and Classification System (HFCAS) framework will be used to elaborate on the findings. This framework is derived in cooperation with the Federal Aviation Authority (FAA) to examine underlying human causal factors in aviation incidents and accidents. It classifies unsafe acts and preconditions for unsafe acts. Adverse mental and physiological states are recognized as unsafe aircrew conditions. Therefore, detection of those states is required to allow for shift to a safe cognitive state.

1.3 Outline of the thesis

This thesis continues with a literature review of related work. Thereafter, the experimental set-up is discussed in Chapter 3. This includes the description of the data, emphasized on the different cognitive states and physiological features. Furthermore, the proposed models are elaborated on in combination with the design procedure. Chapter 4 presents the results and validity of the models. Finally, the discussion and conclusion is provided in respectively Chapter 5 and Chapter 6. The software packages used in this thesis, can be found in Appendix A and will not be explained in detail.

2. Related Work

This thesis builds on research conducted by the National Aeronautics and Space Administration (NASA) on attention management in commercial aviation (Harrivel et al. 2016). The research makes use of a similar dataset as this thesis. However, in contrast to this thesis, NASA also possesses qualitative data gathered during questionnaires conducted after the experiments. Harrivel et al. successfully apply Gradient Boosting, Deep Neural Network and Random Forest methods to classify the pilot's cognitive state and use the means of the three independent results for the overall performance. The sequel of this research has been published in January 2017, and indicates the effort of NASA to improve their own research and recommendations (Harrivel et al. 2017). Finally, a second sequel was published in January 2019 *(Training for Airplane State Awareness using Biofeedback)*. However, this has been withdrawn on the request of the authors. The reason for this is unknown, however, it does indicate the relevance and possibly the quick development within this field. This is corroborated by Kaggle's competition.

NASA's first paper contains research in which non-flight related experiments were performed to simulate cognitive states that can occur during flights. In the subsequent investigation, Harrivel et al. let their participants perform experiments in a motion based simulator for which exercises were defined to simulate the different cognitive states (Stephens et al. 2017). Furthermore, they focus on deriving additional features from one sensing modality. This has for example been accomplished by summarizing statistics over time and by frequency filtering. In total, 1810 features contributed in the prediction of cognitive states. The final findings comprise a prediction accuracy between 0.50 and 0.78 for each individual pilot. It should be noted that this investigation distinguishes seven cognitive states instead of four (additional are high workload, low workload and confirmation bias). Only the average of the applied methods' performance is shared, not the results for each prediction method separately. The qualitative results indicate that most of the pilots (21 pilots out of 24) share none to minimal concerns about performance limitations because of obstruction by equipment.

As well as focusing on the models, it is important to investigate the impact of variables and differences in performance of each cognitive state prediction. Besides the cognitive state classification, researching the cognitive state change (binary observation) might reveal potential capabilities which can be combined with a model capable of predicting state. NASA recommends an emphasis on participant dependent performance. Hence, this research discusses performance of the proposed models for each pilot separately. Due to computational limitations, in contrast to NASA, there will be less focus on overall performance.

2.1 Contribution to aviation safety

The impact of pilot's cognitive prediction on aviation safety is stated by the Commercial Aviation Safety Team (2014) (CAST). One of the Safety Enhancements is named: 'Airplane State Awareness - Training for Attention Management' and focuses on the limitations of human performance within aviation. In 16 out of 18 accidents or incidents, where loss of aircraft control was experienced, issues with flight crew attention were involved. Therefore CAST urges the need for research in attention management by government, industry and academia. As a consequence, NASA published the first

research in this field in 2016. Therefore, minor research has been performed in cognitive state prediction for pilots. According to CAST, potential obstacles for applications in cognitive state prediction can be cost-effectiveness and operationality of practical measures.
Shappell and Wiegmann (2000), fulfilled cooperative research between the Federal Aviation Authority and the University of Illinois. They state, that 70 to 80 percent of the civil and military accidents implicate human error. Additionally, they distinguish nine conditions, which account for an adverse mental state. To increase aviation safety, those should be prevented from happening. In this thesis, three cognitive states and a baseline state are mentioned.
A literature review by Borghini et al. (2014), reveals the relation between physiological characteristics and high mental workload, mental fatigue and drowsiness. Despite the possibilities to detect mental states by physiological measurements, there is only partial common ground with the previous mentioned research by NASA. Borghini et al. focus on research in single state classification, as well for different cognitive states. The results they provide, indicate an increase or decrease of physiological characteristics when entering a cognitive state. The relative contribution of variables to cognitive state prediction indicate possibilities for EEG, heart rate and eye blink rate measurements. This relates to SQ1. Within the research by Borghini et al. (2014) there is referred to the Turkish Airlines accident at Amsterdam Schiphol Airport in 2009. During this fatal accident, the flight crew was not able to recognize the cause of the auto throttle system reducing the thrust of the engines (namely, the malfunction of the radio altimeter) (van Vollenhoven 2010). Within a crucial time span, this moment of seemingly channelized attention, led to an aircraft state from which the crew was unable to recover. It might be difficult to predict in what sense a cognitive state notification could have prevented this accident from happening. Note that the awareness of being in a cognitive state does not solve the problem. The solution however, lies in appointing the adverse mental states to the pilot, such as complacency, overconfidence and misplaced motivation (Shappell and Wiegmann 2000). If a dangerous situation arises, and an algorithm would be able to assign such mental states to the flight crew, they cannot ignore the unsafe condition. Therefore, improvement within the field of cognitive state prediction holds a component to further increase aviation safety.

3. Experimental Set-up

The experimental set-up is explained following a description of the dataset, the feature engineering methods and the prediction models.

3.1 Dataset

The dataset, which is used to predict pilots' cognitive states, is provided by Kaggle and contains approximately six gigabytes of data (Kaggle 2018). Herein lies a challenge because the data include millions of observations and are raw, which therefore contain noise and artefacts. The physiological data of nine flight crews, in total eighteen pilots, are gathered. The crews are always situated as a couple in seat 0 (left/captain) and seat 1 (right/first officer), also when performing non-flying experiments. All variables are measured at a frequency of 256 Hz. However, due to computational limitations, the training data are sampled to derive sixteen observations within each second. The sampling has been performed to safe time during the training procedure, after the data visualization and data processing. As a consequence of the complex characteristics of the dataset, this thesis stresses the importance of feature engineering.
The training dataset comprises activity measurements which are categorized into four cognitive states. Using this data the model will be trained and validated, after which the test data, a line oriented flight training (LOFT), should be tested for the classification capabilities of the model. Because at the time of writing, the actual cognitive sates related to the LOFT have not been released, the model will be tested on the data of two flight crew couples which are separated from the training data. The training set comprises crew 1, 2, 5, 6 and 8, the validation data crew 3 and 7, and the test set crew 4 and 9. The task is to detect the cognitive states according to physiological measurements. Each observation is related to one cognitive state.

Table 1: Cognitive States

Cognitive state	**Abbr.**	**Description**
Baseline	BL	No specific state occurs.
Channelized attention	CA	The state of being focused on one task.
Diverted attention	DA	The state of diverting one's attention by decision making through action or thought.
Startle/Surprise	SS	The state of experiencing rush and adrenaline by observing abrupt change.

3.1.1 Cognitive states. During three experiments for each crew, the data are gathered to train a model which should be capable of predicting the cognitive states during flight operations. In this research four cognitive states are distinguished, namely, channelized attention (CA), diverted attention (DA) and startle/surprise (SS). Finally, a baseline (BL) record is provided within each experiment. As a consequence, only two states can occur within each experiment, namely, the cognitive state which is triggered and the baseline state. Table 1 indicates the distinguished cognitive states. A visualization of the training data distribution over time, is presented in Figure 1.

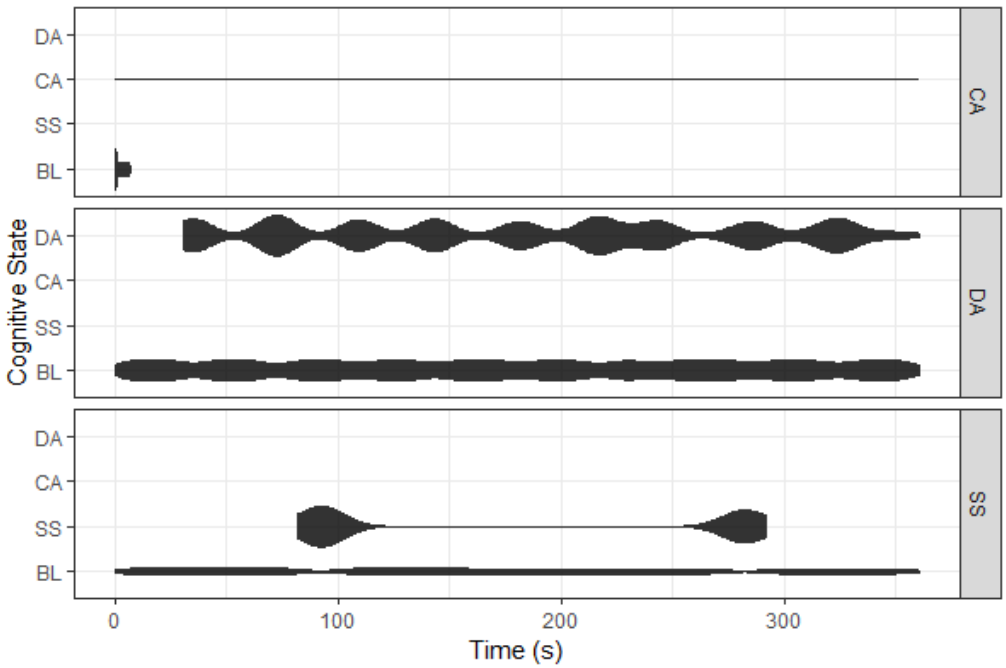
Throughout the three experiments the following is observed for the flight crews within the training data:

1. All experiments take approximately 6 minutes. For each pilot, three experiments are performed which result in roughly 275,000 observations. The total training data comprise 2,649,052 observations before sampling each sixteenth observation.
2. The BL (58%) and CA (34%) state are the most frequent. The DA (5%) and SS (3%) state are the least frequent.
3. In the CA experiment, the CA state is maintained after a short period of BL (a maximum of ten seconds).
4. The SS and DA state occur during multiple periods in its experiment; approximately two periods (SS state) and eight periods (DA) for each pilot.

In contrast to NASA's research, the experiment strategy is only described in a relatively general manner by Kaggle. However, in the upcoming paragraphs the cognitive states are explained on the basis of relevant literature.

Channelized Attention. CA is the state of being focused on one task to the exclusion of all others. During the experiment, this has been achieved by having the pilots perform an engaging puzzle task. As indicated in Figure 1, the pilots stay within the CA state for approximately 6 minutes. This cognitive state involves the ignoring of other tasks and therefore could entail dangerous situations. According to Cheung (1998), channelized attention is "a pilot's attempt to perform a demanding or unfamiliar task, which allows his attention to be confined to one aspect of the task. He/She therefore fails to make optimum use of information about the aircraft orientation." Warning a flight crew about

Figure 1: Distribution of cognitive states for all training data pilots during the three experiments.



reaching a CA state, could enable them to consciously approach a certain situation without being channelized to one aspect.

Diverted Attention. DA is the state of having one's attention diverted by actions or thought processes associated with a decision. During the experiments, the pilots performed a display monitoring task which was periodically interrupted by a mathematical question. In flight operations, diverted attention can lead to dangerous situations which result, for example, in late response and inadequate acting (Cheung 1998). Fatigue could cause being in a DA cognitive state. This indicates the relevance of detecting this mental state.

Startle/Surprise. During the performed experiments, SS is simulated by having the participants watch movie clips with jump scares. Jump scares are experienced by abrupt change, in this case visually and auditory. As a consequence, the participants notice a feeling of rush and adrenaline. According to Rivera et al. (2014), the impact of experiencing SS can have a negative impact on flight safety because of its distracting and interrupting nature. After a 'high intensity stimulus', recovery time can take up to 60 seconds, in which flight performance is reduced (Thackray and Touchstone 1969). Detection of SS is therefore relevant for aviation safety.

3.1.2 Variables. The dataset consists of four groups of variables indicated in Table 2.

Group 1: Electrocardiography. Electrocardiography measurements are used to create an electrocardiogram (ECG). The measurements consist of a 3-point montage and are spread over a wide range of voltages as can be seen in Figure 2a. Each pilot has their measurements normally distributed around a personal average. ECG data can be used to derive multiple heart related characteristics, of which the most known are heart rate (Hr) and heart rate variability. Both are indicators of stress, arousal and mental health (Kim et al. 2018; Kimhy et al. 2010). To enable generalisability, adequate predicting and to derive well-understood variables, it is desirable to retrieve the heart rate from the raw ECG data. By detecting the peak interval within the desirable frequency range the heart

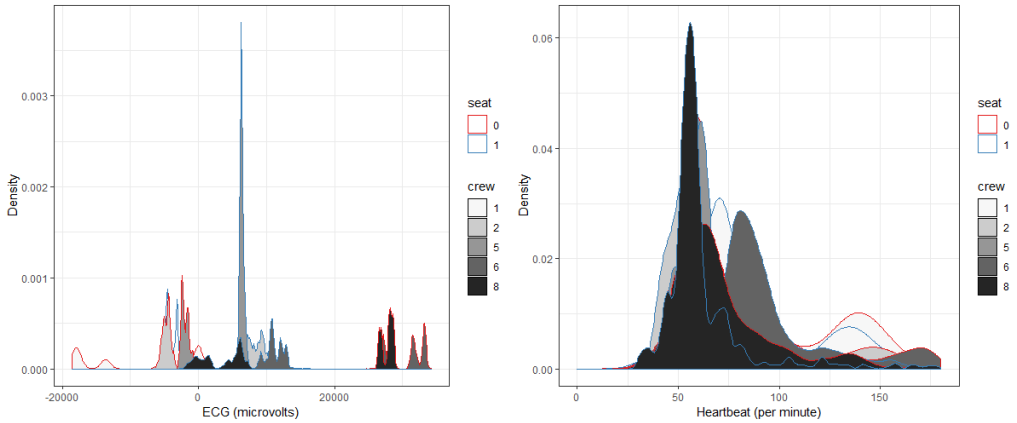
Table 2: The four groups of physiological measurements.

Variable	Abbr.	Unit	Description
Electrocardiogram	ECG	microvolt (µV)	Heart monitoring data. Mainly used to derive the heart beat frequency and variability.
Electroencephalogram	EEG	microvolt (µV)	Monitors the electric activity of the brains which are presented as brain waves.
Galvanic Skin Response	GSR	microvolt (µV)	Refers to the changes in sweat gland activity that are reflective of the intensity of our emotional state, otherwise known as emotional arousal.
Respiration	R	microvolt (µV)	The action of breathing, a measure of the rise and fall of the chest.

rate is calculated. The ECG data are expected to be relevant in between 0.6 and 4 Hz (equal to 36 and 240 heart beats per minute). This has been performed trough frequency domain analysis. Because the ECG processing is highly related to feature engineering, this is further elaborated on in section 3.2.1. The heart rate distribution is visualized in Figure 2b. Variability in heart rate is derived though the sliding window method (section 3.2.2).

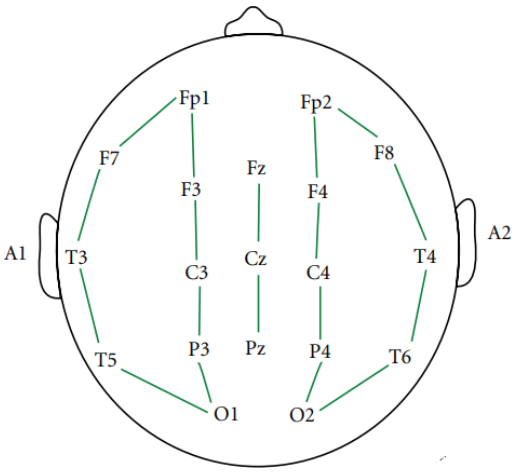
Figure 2: The ECG data prior and after processing, subdivided per crew and seat.

(a) The raw ECG data. (b) The derived heart rate data.



Group 2: Electroencephalogram. Electroencephalogram (EEG) data describe the brain activity of the pilots with the use of twenty electrodes. Those electrodes are likely to detect every local peak in brain activity (Jasper 1958). Unfortunately, no further information is provided by Kaggle about the equipment, nor the montage of the sensors. Several methods of montage are described by Fraga et al. (2011), and tested for support vector machine and logistic regression applications in the diagnosis of Alzheimer disease. The Longitudinal Bipolar method is the second preferred montage method. It is applied in this thesis, simply because all individual EEG sensor data are available for this method. The most preferred method is Bipolar Counterpart, however, by applying this method the data of five sensors would be ignored. Note that finding an appropriate method after the data are gathered is not optimal. This is supported by the fact that the data of an additional sensor are presented within the dataset, namely sensor Poz. To enable generalisability to other pilots, relative EEG values are preferred. This is achieved by calculating voltage differences between two sensors. The sensor connections are visualized in Figure 3. The naming of the sensor is related to the location of the sensor on the head. Sensor Poz is located in between Pz, O1 and O2 (Wild 2007). To enable the use of this sensor's data, Poz is linked to sensor Pz. The voltage differences are calculated from the front of the head to the rear (Maitland 2019). However, different linking directions might function as well and could influence the results on top of the decision to apply the Longitudinal Bipolar method. Nineteen features are derived by calculating the voltage differences between sensors, namely:
Fp1 - F7, F7 - T3, T3 - T5, T5 - O1, Fp1 - F3, F3 - C3, C3 - P3, P3 - O1, Fz - Cz, Cz - Pz, Pz - Poz, Fp2 - F8, F8 - T4, T4 - T6, T6 - O2, Fp2 - F4, F4 - C4, C4 - P4 and P4 - O2.

Figure 3: Longitudinal Bipolar EEG sensor montage (Fraga et al. 2011).



In the result section of this thesis, these features are presented in small letters and connected with an underscore instead of a hyphen sign (eg. fp1_f7). No further data preprocessing has been applied to the EEG data, therefore power grid interferences can be expected. However, this is considered to be negligible because of the use of relative values. The noise is diminished by subtracting the sensor measurements. The noise caused by power grid interference is expected to be equal for all sensors per observation. In EEG research, spectral peaks are often studied to observe the brain state of the patient, such as the Theta band (4-7 Hz), which represents drowsiness, the Alpha band (7-13 Hz), which indicates relaxation and the Beta band (13-35 Hz), which reveals focus (Roohi-Azizi et al. 2017). This is highly related to the cognitive states which are described above. Due to a large amount of sensors, the features are not filtered for frequency. However, a sliding window analysis has been applied to the EEG sensor differences (3.2.2).

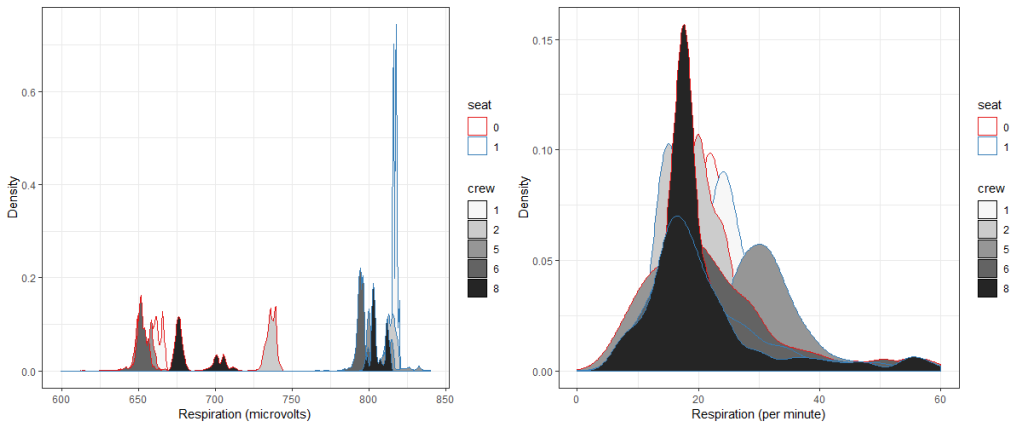
Group 3: Galvanic Skin Response. The Galvanic Skin Response (GSR) describes the electrodermal activity, also known as sweat gland permeability. This is measured by means of skin resistance to small electrical currents (Critchley and Nagai 2013). According to Fernandes et al. (2015), GSR supports the determination of stress and complements the detective capability of the ECG and respiration data. It should be noted that a controversy against GSR exists, because of its unscientific nature (Novella 2015). According to critics, GSR measures merely sweat instead of detecting stress and psychological state. It solely measures skin conductivity, and therefore no distinction can be made concerning the cause of electrodermal activity (e.g. mental stress, anxiety, startle and fear). However, it should be noted that during the research this argument might be applicable to multiple physiological variables.
GSR is not represented as a sinusoidal signal for which the peak interval is relevant. Therefore no frequency filtering is applied to the data. Processing of the data has been performed according to a sliding window analysis (section 3.2.2).

Group 4: Respiration. The respiration data represent the rise and fall of the chest by measuring muscle activation. One's emotional state influences the respiratory characteristics which is therefore an appropriate indicator to identify cognitive states (Feleky 1916;

Homma and Masaoka 2008). Respiratory rate increases as a result of anxiety or stress. As can be seen in Figure 4a, each pilot has their measurement values normally distributed. Remarkably, the data acquired from pilots in the left seat have a significantly higher average, compared to the right seat. This could be due to equipment characteristics.

Figure 4: The Respiration data prior and after processing subdivided per crew and seat.

(a) The raw respiration data. (b) The derived respiration rate data.



Because respiration data are typically shaped sinusoidally, the interval between peak observations might reveal predictive capabilities. Retrieving the respiratory rate (Rr) would also diminish the effect of unequally distributed raw data. After applying noise filtering and locating the peak observations, the respiratory rate is derived. This has been performed through frequency domain filtering, which is further explained in section 3.2.1. The processed respiration data distribution is visualized in Figure 4b.

3.2 Design & Procedure

After the variable analysis, features are extracted from the variables by the application of a frequency domain analysis and sliding window strategy.

3.2.1 Frequency domain analysis. A frequency domain analysis is applied to engineer new features from the complex ECG and respiration data. From the frequency domain filtered signals, the heart rate and respiratory rate are derived. The specific frequency domain which is investigated relates to the external factors of the measured signal. For example, influences of electricity waves are considered noise, whilst the influence of respiration to the ECG data might hold interesting features.

The frequency domain analysis has been performed by a simple low pass filter, which should be seen as a blockage of all frequencies above a set value (Bogner, Constantinides, and Yuen 2008). The filtering is performed by the adjustment of observations according to a smoothing factor (α) to the input signal and to the previous filtered observation (Equation 1). α is based on the relation between the interval of the observed data (Equation 3) and the recursive coefficient factor, depending on the

frequency to filter (Equation 4). Filters which use a recurrence relation are also called Infinite Impulse Response filters. This means that values are defined as a function of the preceding observation, indicated by $Output_{i-1}$.

$$Ouput_i = \alpha\ Input_i + (1 - \alpha)\ Ouput_{i-1} \tag{1}$$

$$\alpha = \frac{\Delta t}{\Delta t + RC} \tag{2}$$

$$\Delta t = \frac{1}{f_{sampling}} \tag{3}$$

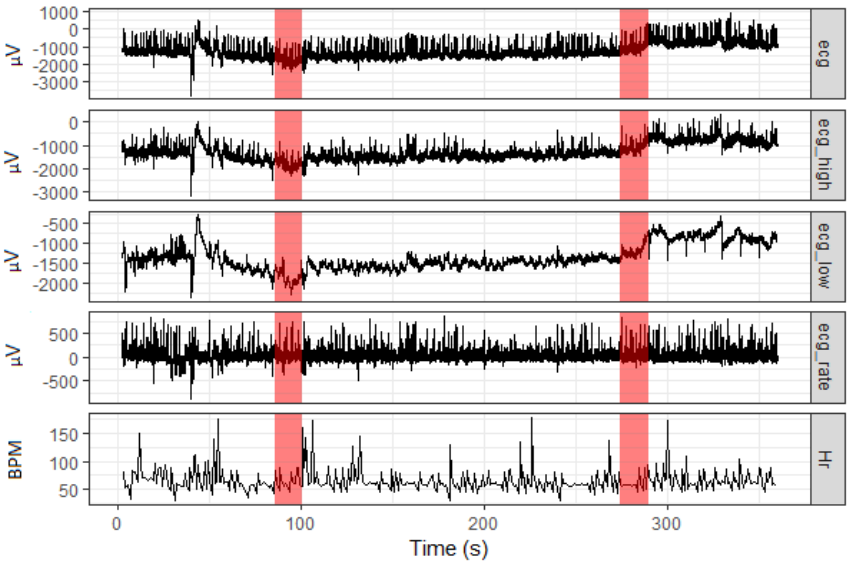
$$RC = \frac{1}{2\pi f_{filter}} \tag{4}$$

To diminish the effect of frequencies above a certain level, the f_{filter} is set. The recursive coefficient and the smoothing factor (α) depend on the filtered frequency only, because the Δt remains equal (considering a stable sampling frequency). The lower the frequency to let pass, the more the filtered signal is defined according to its preceding filtered output (because of a larger RC and smaller α). This is also referred to as exponential smoothing properties.

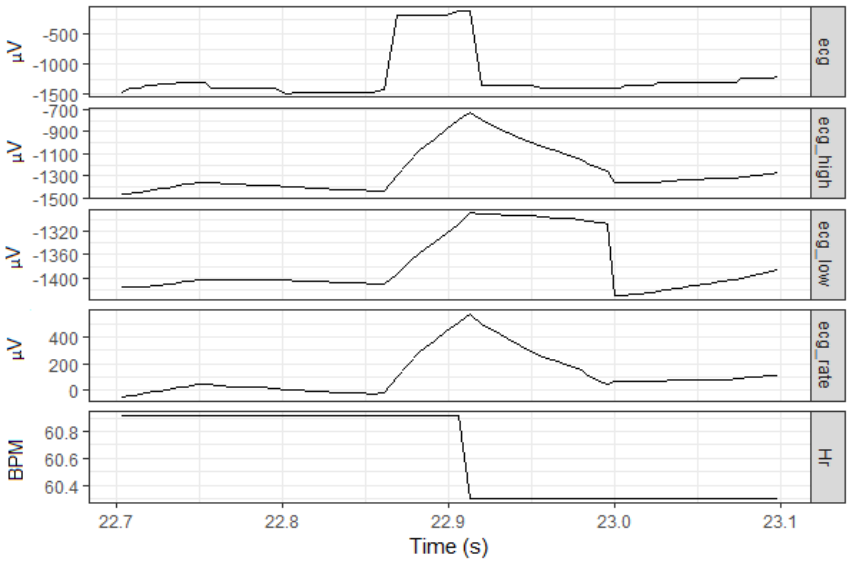
The derivation of the heart rate from the raw ECG data is presented in Figure 5a. This data represent one pilot during the SS experiment. The baseline state is interrupted by the SS cognitive state, indicated by the red bars. As can be seen in the top row, this specific pilot, during this specific SS experiment, shows noise in the ECG measurement. To filter out the low frequencies, a threshold is set at 0.6 Hz. This matches a 36 beats per minute heart rate. Lower heart rate values are not expected and additionally, this is an appropriate filter to detect respiration influences. The high frequency filtering typically rules out the higher frequencies from muscle activity and external electricity waves. This is set at 4 Hz (heart rate of 240 beats per minute). The low frequency filter data are subtracted from the high frequency filter data (ecg_rate = ecg_high - ecg_low). This leaves the ECG data which can be used to detect the heart rate. The heart rate is derived by detecting the interval of voltage peaks for the ecg_rate time series. To deal with missing values, the last observation carried forward method is applied to fill the interval between heart beat observations. In Figure 5b a more detailed visualization of one ECG peak is shown. Compared to the (raw) ECG data in the top row, clearly a more distinguishable peak can be detected in the clean ECG data (ecg_rate). The moment a peak is detected, the new ECG peak interval results in a change in heart rate (unless the peak interval is equal to its previous interval). Note that the y-axis voltage range differs for each (filtered) ECG plot, and heart rate is presented in beats per minute.
Similar to the ECG data processing, a frequency window is determined to retrieve the desirable respiration data. Because respiration normally occurs at a rate of 8 to 35 per minute, the frequency window is expected to be relevant between approximately 0.1 and 0.6 Hz. The results after filtering the appropriate frequencies are shown in Figure 1 in Appendix B. The signal of frequencies below 0.1 Hz is subtracted from the signal of frequencies below 0.6 Hz. This leaves the data of the desirable respiratory rates

Figure 5: Retrieving the heart rate from ECG data, visualized for one pilot during a SS experiment.

(a) Display of the full experiment



(b) Display of 0.4 seconds within the experiment.



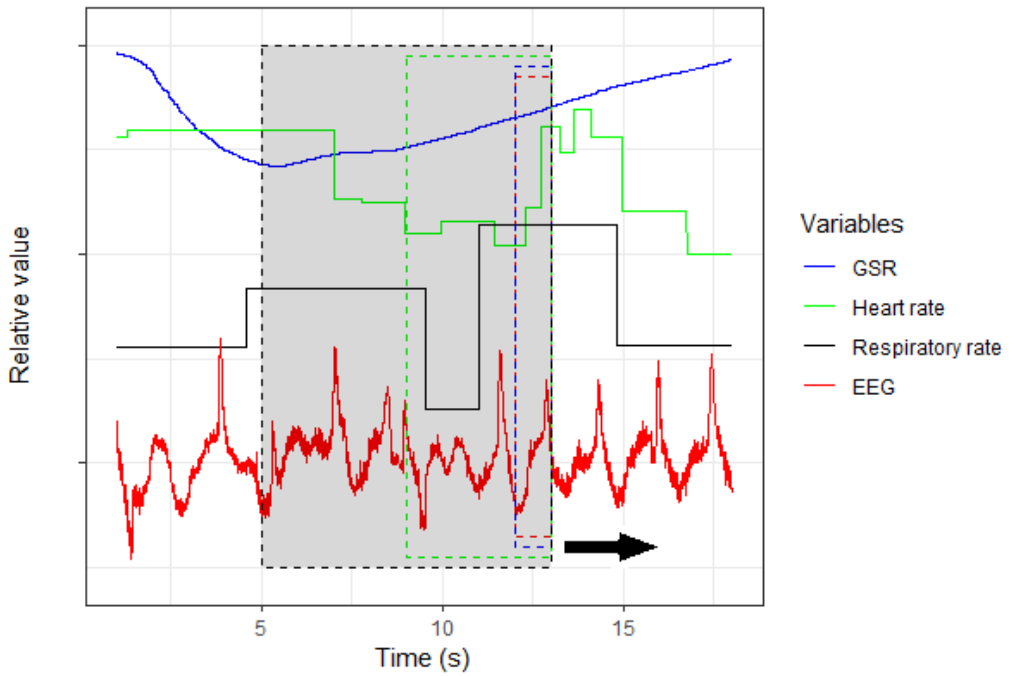
(r_clean = r_high - r_low). Peak detection and determination of the interval results in the respiratory rate (Rr), indicated in the bottom bar. The last observation carried forward method is applied to fill the interval between respiration observations.

3.2.2 Sliding window analysis. To retrieve temporal patterns out of the dataset, the sliding window approach is applied. This is an appropriate method to observe and detect trends in data, presented as time series (Kontaki, Papadopoulos, and Manolopoulos 2005). To enable the use of standard classification models, it is required to retrieve temporal features out of the dataset. The classification models do not consider the sequence of observations within time series data. Therefore features which do not depend on time are required.

Table 3 indicates which features are derived from the variables over a specified time window. Figure 6 visualizes the observed data for the first seconds of one pilot performing the DA experiment. The colour of the variable corresponds to the dashed line of its related time window. Only the Fp1 - F3 feature is indicated in the plot as EEG input, to prevent from indistinctiveness. The y-axis contains no values because the data represent a different scale for each variable. The window slides over the visualized data and describes the trend along with the observed cognitive state. These observations form a new dataset. The temporal trend can be described according to various data characteristics within the sliding window. In this thesis, patterns in data are distinguished according to the following calculations:

1. Mean: The mean function is applied to smoothen values within the time frame and derive long term trends (affix to feature: _mean).

Figure 6: Sliding window time analysis. The dashed boxes indicate the related window size for each variable.



2. First derivative: This function defines the average change of the variable within its time window by subtracting the first observation from the last observation (affix to feature: _der).
3. Standard deviation: This derives variability of all observed values within the time window (affix to feature: _SD).
4. Maximum - minimum value: Detects the difference of the most outlying observations within each time window (affix to feature: _MM).

To allow for a wide variety of patterns being detected, the feature engineering is hardly limited, beyond computational capacity. This is supported by the research of Harrivel et al. (2017). They retrieve 1810 features out of a similar dataset as this thesis.
Because the variables hold different temporal patterns, different window sizes are applied. The decision to select a one second time window for the EEG and GSR data is based on visual clues in Figure 6. The EEG and GSR data show fluctuations within each second, which can be represented as a sliding window feature. Because heart rate and respiratory data are less frequently imputed (because of derivation from EEG and respiratory data), the related time windows are enlarged. Temporal patterns for heart rate are analysed within a sliding time window of 4 seconds and respiratory rate for 8 seconds. Due to less imputations of the heart and respiratory rate, the temporal observation holds less information compared to the EEG and GSR measurement in the same time window.
All four sliding window analysis calculations are applied to the GSR data because for this variable the only feature engineering is performed through this method. For the heart rate, respiratory rate and EEG sensor differences (from nineteen connections) there is focused on the variability features because of their fluctuating patterns.

Table 3: Sliding window features

Variable	**Window size**	**Calculations**	**Features**
Heart rate	4 seconds	Maximum Hr - minimum Hr Mean Hr	2
EEG sensor differences	1 second	Maximum - minimum value Standard deviation	19 x 2
GSR	1 second	Maximum - minimum value Standard deviation Mean value First derivative	4
Respiratory rate	8 seconds	Maximum Rr - minimum Rr Mean Rr	2
		Total sliding window features	46

3.3 Methods

During this research two classification methods are applied. Gradient Boosting Machine (GBM) is used to classify the cognitive states (task 1), and logistic regression is used for the binary classification of cognitive state change (task 2). The sliding window features and the frequency domain features are tested for their predictive performance. Also the initial ECG, GSR, respiration measurements and the EEG sensor differences are tested. If

only to validate the contribution of the complete set of variables. As a result of this thesis stressing the importance of feature engineering from variables, there was not focused on testing multiple classification methods. Despite the awareness of retrieving more robust results when testing for various methods.
For both tasks, the independent variables, including crew, seat, time and the experiment title, are left out. A benchmark model is trained to validate the use of engineered features in performing task 1. This enables the comparison between a GBM model with and without engineered features.

3.3.1 Task 1: Cognitive state classification. The classification of cognitive states is accomplished by a Gradient Boosting Machine (GBM) method (Friedman 2002). GBM typically applies to classification and regression problems, and benefits in execution speed and accuracy from a randomization procedure by sampling the training data. According to a loss function, the learner is modified and improved with each iteration. GBM is characterized by the use of a pseudo residual, which is the gradient of the loss function. GBM is suitable for this research because of the following. Compared to standard classification tree analysis, GBM is less limited because of combining the multiple trees iteratively (Lawrence et al. 2004). Additionally, GBM is less affected by outliers, inaccurate training data and unbalanced datasets. Finally, GBM is capable of dealing with complex models and the contribution of large amounts of features to the prediction task. In previous research, GBM appeared to be effective in various fields, such as chemistry (Babajide Mustapha et al. 2016), ecology (Moisen et al. 2006) and medical science (Xu et al. 2017). Boosting is generally referred to as the increase of accuracy in existing machine learning methods and therefore not necessarily related to stochastic gradient methods only (E. Schapire 2002).
Whereas the initial GBM model to derive the data's feature importance ran on default settings, the final model's tuning parameters were set according to three parameters. The number of trees and the interaction depth indicate the complexity of the model. The learning rate refers to the adaptive amplitude to the loss function. The parameters settings are presented together with the results in section 4.1. Cross validation is used during the training procedures to derive the predictive probability losses.

3.3.2 Task 2: Cognitive state change detection. To fulfil task 2, a relatively simple logistic regression method is applied. The detection of a cognitive state change is a binary decision. The data are sampled to retrieve one observation for each second. An additional column is added, which indicates whether a change of cognitive state has occurred compared to the previous observation. Only *True* or *False* is presented, the initial state, or subsequent state is not mentioned. The use of multiple predictors allow for logistic regression in binomial data (McDonald 2014). Thereby, logistic regression can be used to derive the predictive capabilities of the features independently. A literature review by Christodoulou et al. (2019), found no evidence that for binary classification within clinical research, machine learning algorithms outperform logistic regression. This substantiates the use of this method.

3.4 Software

The data processing and modification are performed in the *R* environment. The packages consulted are specified in Appendix A. The standard packages, which do not entail characteristics needed for reproduction of the experiments, are left out. Most of the

predictive functions are applied according to the *caret* package manual (Kuhn et al. 2018).

3.5 Evaluation Criteria

The models' performance will be derived according to task specific methods. For the cognitive state classification model (task 1), the F1 score will be consulted. This metric is a trade off between the precision, which indicates the amount of *True Positives* over all the *Positive* predictions, and the recall, which indicates the *True Positives* over all actual *Positives*. This enables the detecting of less frequent cognitive states. For example reviewing the accuracy (which Harrivel et al. (2016) apply), would be inappropriate because of the frequent occurrence of the baseline cognitive state. The model is internally optimized according to the mean F1 metric with the use of cross validation. To enable the comparison with the results from NASA, accuracy is mentioned in the results as well. Feature selection of the GBM model is accomplished by reviewing the cross entropy loss function. This indicates the contribution in predictive probability of each feature to the classification task (Janocha and Czarnecki 2017).

The logistic regression model's performance is reviewed according to the Area Under Curve (AUC) value. This measures the performance for different threshold values for the binary classification prediction probabilities. Similar to task 1, performance in cognitive state change detection could be measured according to the F1 score. However, this does not indicate which probability threshold is used to determine the binary decision. The feature selection for task 2 is based on standard statistical analysis.

4. Results

The results of both performed classification tasks are presented, after which they are discussed in Chapter 5. For the cognitive state classification task, a benchmark to feature engineering is provided. This enables the validation of applying engineered features compared to the raw unprocessed variables.

4.1 Task 1: Cognitive state classification

The task 1 model is trained on a sample of the measurements (each sixteenth observation) from the pilots of crew 1, 2, 5, 6 and 8. The feature importance is tested on crew 3 and 7. According to the most contributing features, the adapted model is trained and validated on the previous mentioned flight crews. Finally, the performance is tested on a separated test set, which include the pilots of crew 4 and 9.

Figure 7: Top ten predictive contributors according a cross entropy (ce) loss function.

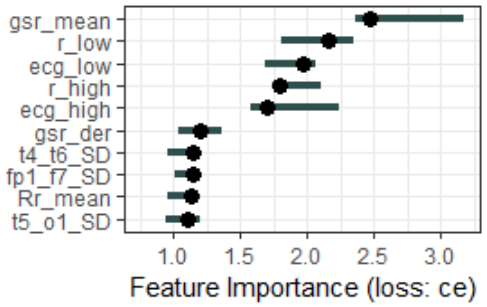


Table 4: Feature explanation of Figure 7

Feature	Explanation
gsr_mean	The sliding window (1 second) mean of the GSR .
r_low	The low pass, low frequency (0.1 Hz) filter of the respiration data.
ecg_low	The low pass, low frequency (0.6 Hz) filter of the ECG data.
r_high	The low pass, high frequency (0.6 Hz) filter of the respiration data.
ecg_high	The low pass, high frequency (4 Hz) filter of the ECG data.
gsr_der	The sliding window (1 second) derivative of the GSR .
t4_t6_SD	The sliding window (1 second) standard deviation of the difference in voltage between EEG sensors T4 and T6. .
fp1_f7_SD	The sliding window (1 second) standard deviation of the difference in voltage between EEG sensors Fp1 and F7.
Rr_mean	The sliding window (8 seconds) mean of the respiratory rate.
t5_o1_SD	The sliding window (1 second) standard deviation of the difference in voltage between EEG sensors T5 and O1.

To eliminate the least contributing features, initially task 1 is performed on default settings. The feature contribution is derived from the validation data, and is visualized in Figure 7. For each feature a short explanation is indicated in Table 4. The top ten out of 73 contributors are indicated, in which a clear distinction can be identified for the

best five predictive features. The feature importance is derived according to a 'leave-one-out' principle. The cross entropy loss, is therefore the highest when leaving out the sliding window mean of the GSR. It should be interpreted as losing the highest level of predictable probability, when leaving out this variable.
The adapted version of the task 1 model predicts the cognitive state event based on the sliding window mean of the GSR (gsr_mean), the 0.1 Hz frequency filter of the respiration (r_low), the 0.6 Hz filter of the ECG (ecg_low), the 0.6 Hz filter of the respiration (r_high) and the 4 Hz frequency filter of the ECG (ecg_high). The frequency domain analysis seems therefore to contribute considerably in the derivation of relevant features. The settings of the parameters of the adapted model are indicated in Table 5. A wide range of numbers of trees and interaction depth is chosen to allow for the research of complex and less complex model settings. Computational restrictions affected the parameter settings, such as the testing of a lower learning rate. The training and validation is performed on 233,804 sampled observations of the five mentioned features. The test dataset contains 1,102,736 observations from flight crew 4 and 9, and is not sampled.

Table 5: GBM settings

Parameter	**Setting**
Number of trees	3, 6, 9
Interaction depth	50 : 250
Learning rate	0.1
Loss metric	Mean F1
Train control	5 fold cross validation

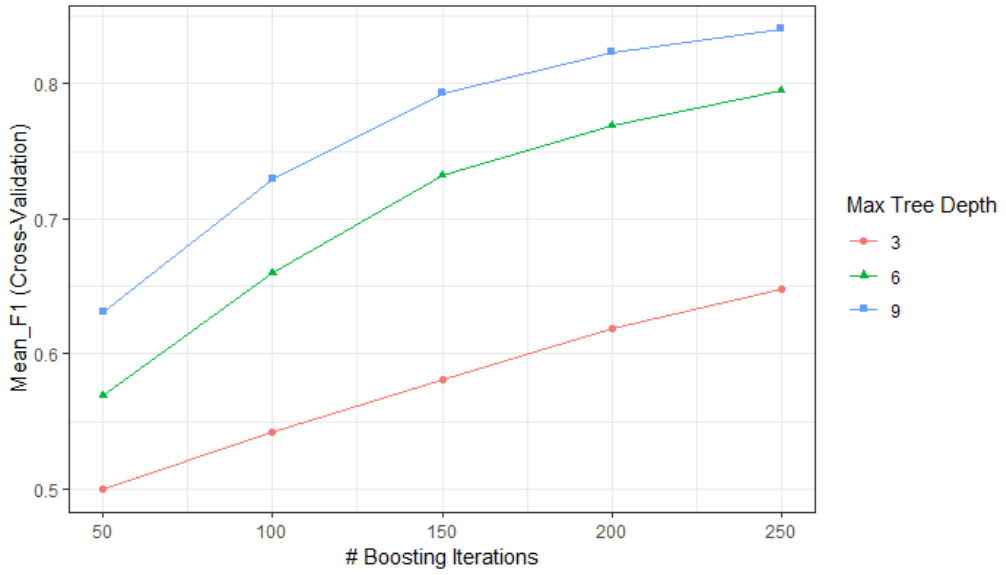
The results of the adapted model for task 1 are presented in Table 6. Additionally, the performance of a benchmark model with respect to the feature engineering is shown. These are retrieved from a model trained, validated and tested according to the same parameters and flight crews. However, this benchmark model uses the initial unprocessed variables of the dataset to predict cognitive state. Consisting of raw data from ECG, GSR, Respiration and twenty EEG sensors.

Table 6: Task 1 performance (F1 score)

	Engineered features		**Unprocessed variables**	
	Training/validation	**Test**	**Training/validation**	**Test**
BL	0.92	0.73	0.89	0.64
SS	0.77	NA	0.56	NA
CA	0.88	0.33	0.88	0.31
DA	0.59	0.04	0.49	0.005
Weighted F1 score	**0.89**	**0.55**	**0.86**	**0.48**

The results contain the predictions during the experiments of all pilots combined. The confusion matrix for both the training/validation and the test dataset are shown in Appendix C. As well for the benchmark model. The F1 score is presented for each separate classification and as a weighted overall performance score. Notice that no SS states are predicted for the test dataset. Especially prediction of the DA and SS cognitive

Figure 8: The optimization function of the GBM.



state score remarkably lower on the test dataset, compared to the training/validation data. The overall F1 score is weighted according the occurrence of each cognitive state, and indicates a 0.34 lower F1 score for the testing data. This suggests overfitting of the training data. The weighted F1 score of the engineered features (0.55) are considerably higher compared to the unprocessed variables (0.48) for the test dataset. Only for detecting the CA state in the training/validation data, feature engineering does not improve the performance.

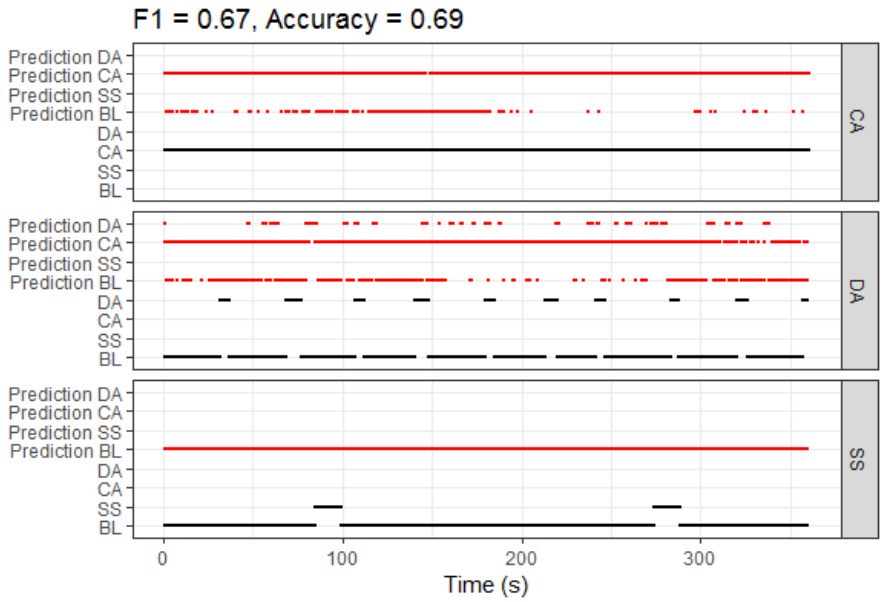
The learning process of the model is visualized in Figure 8. The model has picked the optimal setting according to a 5 fold cross validation. Which results in an interaction depth of 9, and 250 as the optimal number of trees for this model. Due to computational limitations, no more boosting iterations are performed and no 'deeper' models are tested, although the F1 score did not converge to its optimum yet. This indicates no complete retrieval of information from the training data. However, overfitting is already observed when comparing the training and testing F1 score.

Table 7: Inidvidual performance on task 1

Pilot	**F1 Score**	**Accuracy**
Crew 4, left seat	0.44	0.59
Crew 4, right seat	0.44	0.59
Crew 9, left seat	0.67	0.69
Crew 9, right seat	0.45	0.56

4.1.1 Pilot dependent results. Pilot dependent performance has been investigated because of the recommendation by Harrivel et al. (2017). The prediction of the left seat

Figure 9: Cognitive state prediction for the pilot of crew 9, in the left seat.



pilot of crew 9 is visualized in Figure 9. Out of four pilots, this figure represents the best predictive values. The F1 score for this pilot's result is the highest out of four, namely 0.69 (Table 7). The plots of the three remaining pilots in the test set are shown in Appendix D. Note that for predicting solely the BL state, an F1 score of 0.44 is achieved. Three out of four pilots in the test set indicate similar F1 scores (around 0.45). Remarkably, the pilot in the left seat of crew 9, scores considerably higher than the other pilots within the test dataset. This indicates little capacity of the task 1 model to generalise the training data to the four pilots of the test set.
To enable a comparison to the research by NASA, also the accuracies of each pilot dependent prediction is presented. These vary between 0.56 and 0.69.

4.2 Task 2: Cognitive state change detection

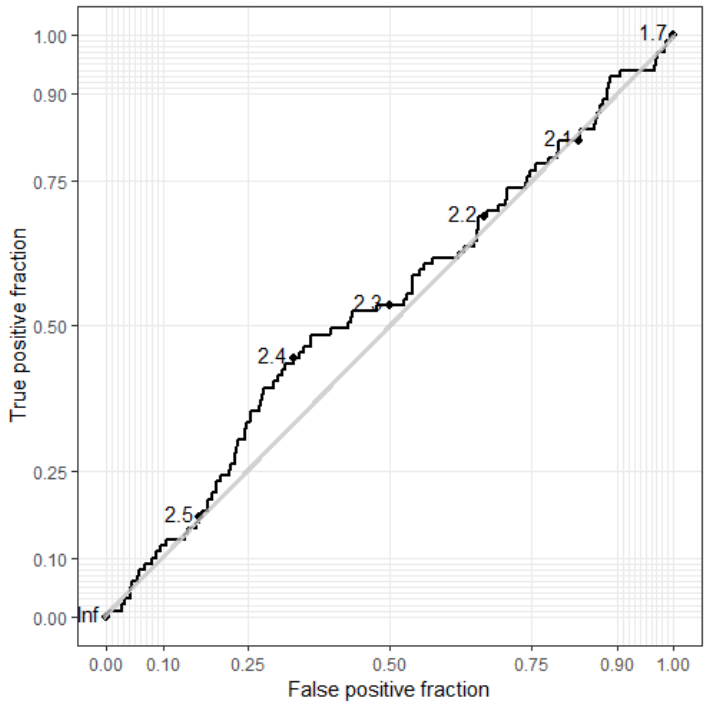
Initially, the task 2 model is trained on a sample of the measurements (each sixteenth observation) from the pilots of crew 1, 2, 3, 5, 6, 7 and 8. The statistical feature contribution is derived from the summary function of this model. According to the significant contributing features, the adapted logistic regression model is trained on the same pilots' measurements. The task 2 performance is tested on a test set, which comprises the pilots of crew 4 and 9.
All variables with a significant impact ($P < 0.05$) on predicting cognitive state change, are indicated in Table 8. Of significant contribution are the heart rate (Hr), the difference between EEG sensor C4 and P4 (c4_p4), the sliding window maximum value minus the minimum value of the difference between EEG sensor C3 and P3 (c3_p3_MM) and the sliding window mean of the GSR (gsr_mean). The logistic regression model derived from those features result in a predictive model which scores an AUC of 0.532 on the test dataset. This represents the performance of both flight crews combined (4 and 9).

Table 8: Logistic regression feature contribution.

Variable	Estimate (10e-2)	Standard Error (10e-2)	P Value
Hr	0.700	0.268	0.009 **
c4_p4	-2.56	1.10	0.019 *
c3_p3_MM	2.50	0.942	0.008 **
gsr_mean	-0.040	0.013	0.002 **

Figure 10 visualizes the ROC curve and indicates the threshold settings for the logistic regression predictions. The best predictive results are achieved with a decision threshold in between 0.02 and 0.03 (the indicated threshold values times 10e-2). The AUC for this binary classification is close to random binary prediction. Therefore it should be remarked, that although the four mentioned features contribute significantly, they contribute in low performance predictions. The performance on task 2 is not discussed for each pilot separately.

Figure 10: ROC plot for logistic Regression with AUC = 0.53. Indicated threshold values times 10e-2.



5. Discussion

The results are discussed according to the research questions.

The first research question investigates whether the cognitive state of a pilot can be predicted from physiological measurements. This thesis affirms the research question up to a certain extent. The model achieves an F1 score of 0.55 on this task, which is fulfilled according to a GBM model. Major differences in performance exist between the training and test dataset. Overfitting of the training data is observed and should be diminished to improve cognitive state classification.

The first sub-question researches the relative contribution of the variables, which is focused on in this thesis. Feature engineering through frequency domain analysis and sliding window analysis, revealed predictive capabilities of the variable groups beyond what it initially measured. According to the frequency domain analysis, the respiration and ECG data were divided into three features. A low pass high frequency filter, a low pass low frequency filter and the difference between those. Determination of the frequency filtered domain, depends on the nature of the variable. For example, the low frequency filter of the ECG data, is set at 0.6 Hz to detect respiratory influences. This feature appears to be considerably contributing to the prediction of task 1 (cognitive state classification). According to a cross entropy loss function, the ECG low filter contains better predictive characteristics than the feature which is designed to derive the respiration data from the respiratory measurements. This is interesting, because the ECG equipment is not intended to retrieve respiratory measurements. Figure 7 indicates that out of the five most contributing features, four were derived according to the frequency domain analysis. Both the high and low frequency filters of the ECG and respiration variables. The highest contribution is achieved by the one second sliding window mean of the GSR. The model which applied the engineered features achieves a 0.07 higher F1 score performance compared to unprocessed data (benchmark). This validates the use of engineered features.
With regards to task 2, detection of cognitive state change, four features appeared to be contributing significantly. Namely, the sliding windows mean of the GSR, heart rate, the EEG sensors difference of C4 and P4, and the sliding window maximum value minus the minimum value of the difference between EEG sensors C3 and P3. It should however be noted, that these features contribute significantly to a low performance prediction.
Frequency domain analysis and sliding window analysis are successful strategies to enable feature engineering. Performance in the prediction task is improved by applying the retrieved features. Feature engineering is hardly limited because of the high amount of possibilities in time series trend detection. Also the frequency domain can be adapted for analysis of each measured frequency. The main remark to the feature engineering in this thesis is the lack of research in frequency band analysis for the EEG data. This was mentioned to be promising, however, because of a large amount of features it is ignored.

The second sub-question considers the relative difficulty in prediction of various cognitive states. The results indicate clear differences in predictability. The results of all pilots within the test dataset are combined. The predictability of the baseline state scores the highest (F1 = 0.73). The SS state has not been predicted during all experiments within the testing data, whilst for the training data an SS F1 score of 0.77 is achieved. The DA state scores low on predictability (F1 = 0.04) in the test dataset,

however this state already indicated lower performance on the training data (F1 = 0.59) compared to the other cognitive states. Prediction of the CA state appeared to be reasonably possible for the left seat pilot of crew 9 only.
The most frequent occurring cognitive states show better performance in predictability. This might be prevented by focusing on the training of SS and DA states separately. For further research, the data could be balanced for the occurrence of cognitive states. However, it was expected that the GBM method is capable of dealing with unbalanced datasets.

The third sub-question considers the pilot dependent performance on task 1. Out of the four pilots in the test dataset, large differences are observed within the F1 score. Three pilots indicate similar performance (F1 = 0.44/0.45) whilst the left seat pilot of crew 9 scores 0.67 on F1 performance. This is a major difference, and therefore the question arises whether a single prediction model would fit a physiological diverse set of pilots. According to the predictive results of the two flight crews within the test dataset, this would not the possible. Either a more generalisable method is needed to enable task 1 by a pilot independent model, or pilot dependent models should be fitted.

The presented quantitative results indicate the possibility as well as the difficulty of predicting cognitive states for pilots. The applicability of the model depends mostly on the predictive performance, which at the moment, is not sufficient for all pilots. A more generalisable manner for predictions is required, which is substantiated by the large differences in F1 scores between the training and test data, and differences in test data results per pilot.
Prediction of cognitive state change does not reveal promising characteristics. The AUC score of 0.532 is close to random prediction of change in state. Now, one second of detection time is used. Allowing for a longer detection time, might contribute to a less complex logistic regression model and more accurate prediction.
Further research should however, focus on complementation of logistic regression models to the more sophisticated machine learning algorithms. Herein lies an interesting observation, namely, detection of the specific cognitive state occurs multiple times within the appropriate experiment according to the GBM model, as indicated in Figure 9. However, not at the exact moment in time. Therefore, a combination of classification models could entail performance improvement.

5.1 Related Work

Several times throughout this thesis, a comparison is made to NASA's complex predictive models. Their pilot dependent accuracy scores lie in between 0.50 and 0.78, which is similar to the results presented in this thesis (accuracy of 0.56 to 0.69) (Harrivel et al. 2017). The necessity of more complex models, and the use of large amounts of features is therefore not supported. Although the computational resources of NASA are expected to be larger compared to this research, the same problems arise during their research. Generalisability of the data to a physiologically diverse set of pilots, appears to be challenging. This could be solved by tailoring a model for each pilot. Obviously, this would come with a tremendous amount of labour. Unfortunately, no results are shared for each model independently, which makes it impossible to validate the performance of the GBM method specifically.
Research by Shappell and Wiegmann (2000), presents the Human Factors Analysis and Classification System (HFCAS), which indicates the various classifications of unsafe acts

within aviation, and the possible pre-conditions for those unsafe acts. Especially the adverse mental and physiological states require attention. During this thesis only three potentially unsafe cognitive states are presented. In the HFCAS, nine adverse mental states are presented, of which channelized attention is one of them. Elaboration beyond the current researched cognitive states (CA, DA and SS) is therefore required.

5.2 Further research

Improvement of task 1 and 2 performance might be achieved by normalizing the trained data and the incubation of calibration procedures during the experiments. NASA already has more data at their disposal in their latest research, and also this thesis could benefit from more pilots performing cognitive states experiments. It should however be noted that this thesis makes use of sampled data. And therefore not utilises the full capacity of the data.
Although mentioned in this thesis, also a frequency domain analysis should be applied to the EEG data to detect frequency bands related to cognitive states. Unfortunately this is left out in this research. Additionally, less complex models could solve the large performance discrepancy between the training and testing data. Overfitting of the training data is observed, and future research should prevent this by researching more 'shallow' models. The research could be broadened by testing different kinds of predictive models.
Finally, fully understanding of the data and determination of cognitive states depending features shape the recommendation for similar research in the future. Harrivel et al. (2017), already researched the possibilities of applying different measurements, such as eye tracking. Possibly, the currently researched physiological characteristics of a pilot, do not contain the key to enable higher performance in cognitive state prediction.

6. Conclusion

The conclusions drawn from this thesis are presented according to five key findings.

Key finding 1. It can be concluded that the applied feature engineering enabled the prediction of cognitive states. The sliding window analysis, and especially the frequency domain analysis, result in retrieval of contributing features out of the complex dataset. Feature engineering on GSR, ECG and respiration data contributes the most. In comparison to cognitive state classification according to unprocessed features, engineered features perform on average 0.07 higher on F1 score.

Key finding 2. Task 1, cognitive state classification, can be achieved according to an F1 score of 0.55. This does not show improved performance with regards to previous research. However, in this thesis, this was not the point of focus. The accuracy of the model lies in between 0.56 and 0.69 for each pilot within the test dataset. This is in line with the performance of NASA's research.

Key finding 3. Cognitive state change is difficult to detect according to the retrieved features from the physiological measurements. An AUC of 0.53 has been achieved.

Key finding 4. Major differences exist in detectability of cognitive state. The most challenging cognitive states to classify are SS and DA. The SS prediction did not take place during all experiments of the test dataset. The F1 score of 0.73 for the baseline state in the test dataset, indicates relatively good detectability

Key finding 5. Task 1 performance is highly influential per pilot. Out of four pilots within the test dataset, one revealed obvious better classification characteristics compared to the others. The left seat pilot of crew 9 scored at least 48 percent higher on F1 score than other pilots. The cognitive state prediction depends on individual characteristics which require more attention in further research.

The model reveals potential success, however, for further application, improvement is necessary. Cognitive state change prediction might hold a feature which could improve the performance of the cognitive state prediction model. However, its current performance is not sufficient to contribute. The relevance of this area of research has been proven by conducting literature research, and is supported by the fact that NASA and the FAA strive to improve aviation safety, among other things, by prediction of pilots' cognitive states.
This thesis contributes in the search for decent feature engineering methods, to retrieve information out of the complex dataset. The results of the research expose the difficulty in generalisability of the model. Physiological characteristics depend on the individual. Reaction to the different cognitive states vary per person as well. The complex physiological data contain a key to the pilots' cognitive state and therewith the aircraft state. In combination with modern AI technology it has potential capability to increase aviation safety, which should never be omitted.

References

Babajide Mustapha, Ismail, Faisal Saeed, Ismail Babajide Mustapha, and Faisal Saeed. 2016. Bioactive Molecule Prediction Using Extreme Gradient Boosting. *Molecules*, 21(8):983.

Bogner, R. E., A. G. Constantinides, and C. K. Yuen. 2008. Introduction to Digital Filtering. *IEEE Transactions on Systems, Man, and Cybernetics*, 8(5):417–417.

Booz Allen Hamilton. 2019. About us.

Borghini, Gianluca, Laura Astolfi, Giovanni Vecchiato, Donatella Mattia, and Fabio Babiloni. 2014. Neuroscience and Biobehavioral Reviews Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload , fatigue and drowsiness. *Neuroscience and Biobehavioral Reviews*, 44:58–75.

Cheung, B. 1998. Recommendations to Enhance Spatial Disorientation Training for the Canadian Forces. (98).

Christodoulou, Evangelia, Jie Ma, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, and Ben Van Calster. 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110:12–22.

Commercial Aviation Safety Team. 2014. Airplane State Awareness.

Critchley, Hugo and Yoko Nagai. 2013. *Electrodermal Activity (EDA)*. Springer New York, New York, NY.

E. Schapire, Robert. 2002. The boosting approach to machine learning: An overview. *Nonlin. Estimat. Classif. Lect. Notes Stat*, 171.

Feleky, Antoinette. 1916. The influence of the emotions on respiration. *Journal of Experimental Psychology*, 1(3):218–241.

Fernandes, Atlee, Rakesh Helawar, R. Lokesh, Tushar Tari, and Ashwini V. Shahapurkar. 2015. Determination of stress using Blood Pressure and Galvanic Skin Response. *2014 International Conference on Communication and Network Technologies, ICCNT 2014*, 2015-March:165–168.

Fraga, F. J., L. R. Trambaiolli, A. C. Lorena, P. A M K Kanda, R. Nitrini, and R. Anghinah. 2011. Does EEG montage influence alzheimer's disease electroclinic diagnosis? *International Journal of Alzheimer's Disease*, 2011.

Friedman, Jerome H. 2002. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4):367–378.

Gosling, Geoffrey D. 1987. Identification of artificial intelligence applications in air traffic control. pages 27–38.

Greenwell, Brandon, Bradley Boehmke, Jay Cunningham, and GBM Developers. 2019. *gbm: Generalized Boosted Regression Models*. R package version 2.1.5.

Harrison, L., P. Saunders, and J. Janowitz. 1994. Artificial intelligence with applictions for aircraft.

Harrivel, Angela R., Charles Liles, Chad L. Stephens, Kyle K. Ellis, Lance J. Prinzel, and Alan T. Pope. 2016. Psychophysiological Sensing and State Classification for Attention Management in Commercial Aviation. pages 1–8.

Harrivel, Angela R, Chad L Stephens, Robert J Milletich, Christina M Heinich, Mary Carolyn Last, J Napoli, Nijo A Abraham, Lawrence J Prinzel, Mark A Motter, and Alan T Pope. 2017. Prediction of Cognitive States during Flight Simulation using Multimodal Psychophysiological Sensing. pages 1–10.

Homma, Ikuo and Yuri Masaoka. 2008. Breathing rhythms and emotions. *Experimental Physiology*, 93(9):1011–1021.

James, Gareth, Daniela Witten, Trevor Hastie, and Rob Tibshirani. 2017. *ISLR: Data for an Introduction to Statistical Learning with Applications in R*. R package version 1.2.

Janocha, Katarzyna and Wojciech Marian Czarnecki. 2017. On Loss Functions for Deep Neural Networks in Classification. pages 1–10.

Jasper, H H. 1958. Report of the Committee on Methods of Clinical Examination in Electroencephalography. Appendix: The Ten Twenty Electrode System of the International Federation. *Electroencephalography and Clinical Neurophysiology*, 10(2):370–375.

Kaggle. 2018. Reducing commercial aviation fatalities.

Kelly, Damien and Merina Efthymiou. 2019. An analysis of human factors in fifty controlled flight into terrain aviation accidents from 2007 to 2017c. *Journal of Safety Research*.

Kim, Hye Geum, Eun Jin Cheon, Dai Seg Bai, Young Hwan Lee, and Bon Hoon Koo. 2018. Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry Investigation*, 15(3):235–245.

Kimhy, David, Philippe Delespaul, Hongshik Ahn, Shengnan Cai, Marina Shikhman, Jeffrey A. Lieberman, Dolores Malaspina, and Richard P. Sloan. 2010. Concurrent measurement of "real-world" stress and arousal in individuals with psychosis: Assessing the feasibility and validity of a novel methodology. *Schizophrenia Bulletin*, 36(6):1131–1139.

Kontaki, Maria, Apostolos N. Papadopoulos, and Yannis Manolopoulos. 2005. Continuous trend-based classification of streaming time series. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3631 LNCS:294–308.

Kuhn, Max, Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan, and Tyler Hunt. 2018. *caret: Classification and Regression Training*. R package version 6.0-81.

Lawrence, Rick, Andrew Bunn, Scott Powell, and Michael Zambon. 2004. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment*, 90(3):331–336.

Maitland, Stu. 2019. Introduction to physiological data.

McDonald, J.H. 2014. *Handbook of Biological Statistics (3rd ed.)*. Sparky House Publishing, Baltimore, Maryland.

Moisen, Gretchen G., Elizabeth A. Freeman, Jock A. Blackard, Tracey S. Frescino, Niklaus E. Zimmermann, and Thomas C. Edwards. 2006. Predicting tree species presence and basal area in Utah: A comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecological Modelling*, 199(2):176–187.

Molnar, Christoph and Bernd Bischl. 2018. iml: An r package for interpretable machine learning. *JOSS*, 3(26):786.

Novella, Steven. 2015. Galvanic skin response pseudoscience.

Obermeyer, Ziad and Ezekiel J Emanuel. 2016. Predicting the Future-Big Data, Machine Learning, and Clinical Medicine HHS Public Access. *N Engl J Med*, 375(13):1216–1219.

Oster, Clinton V, John S Strong, and C Kurt Zorn. 2013. Research in Transportation Economics Analyzing aviation safety: Problems, challenges, opportunities. *Research in Transportation Economics*, 43(1):148–164.

Rich, E. 1983. Artificial intelligence.

Rivera, Javier, Andrew B. Talone, Claas Tido Boesser, Florian Jentsch, and Michelle Yeh. 2014. Startle and surprise on the flight deck: Similarities, differences, and prevalence. *Proceedings of the Human Factors and Ergonomics Society*, 2014-January:1047–1051.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77.

Roohi-Azizi, Mahtab, Leila Azimi, Soomaayeh Heysieattalab, and Meysam Aamidfar. 2017. Changes of the brain's bioelectrical activity in cognition, consciousness, and some mental disorders. *Medical Journal of the Islamic Republic of Iran*, 31(1):307–312.

Shappell, Scott A and Douglas A Wiegmann. 2000. The Human Factors Analysis and Classification.

Skocik, Michael, John Collins, Chloe Callahan-Flintoft, Howard Bowman, and Brad Wyble. 2016. I tried a bunch of things: The dangers of unexpected overfitting in classification. *bioRxiv*.

Stephens, Chad L., Angela R. Harrivel, Lawrence J Prinzel, Ray Comstock, Nijo A Abraham, and Alan T. Pope. 2017. Crew State Monitoring and Line-Oriented Flight Training for Attention Management. *19th International Symposium on Aviation Psychology (ISAP 2017*, pages 1–6.

Thackray, Richard I. and R. Mark Touchstone. 1969. Recovery of motor performance following startle. *Report No. FAA-AM-69-21*.

van Vollenhoven, Pieter. 2010. Dutch Safety Board investigation of Boeing 737-800 crashed during approach, 25 February 2009. page 228.

Wild, Heather A. 2007. Applying Signal Cetection Theory to evoked response potentials for understanding mechanisms of bias and sensitivity in face detection tasks. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, 67(8-B):4737.

Xu, Qian, Yi Xiong, Hao Dai, Kotni Meena Kumari, Qin Xu, Hong Yu Ou, and Dong Qing Wei. 2017. PDC-SGB: Prediction of effective drug combinations using a stochastic gradient boosting algorithm. *Journal of Theoretical Biology*, 417(January):1–7.

Zeileis, Achim and Gabor Grothendieck. 2005. zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, 14(6):1–27.

Appendix A: R Packages

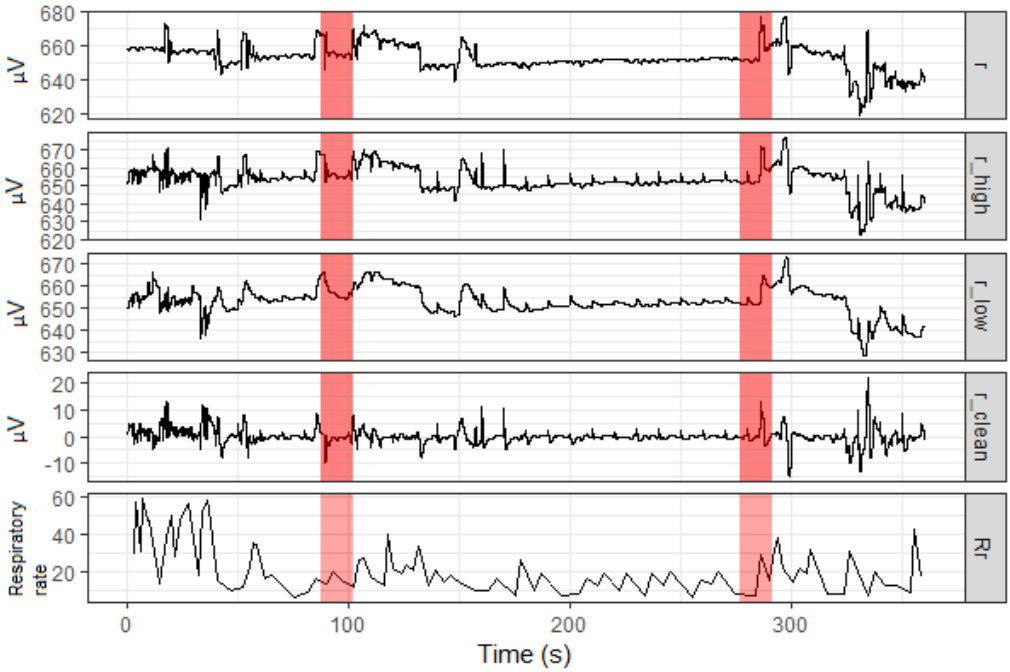
Table 1: R Packages

Title	Describtion	Developers
caret	Training and testing environment for creating predictive classification models.	Kuhn et al. (2018)
GBM	Additional package to enable gradient boosting machine methods in the Caret package.	Greenwell et al. (2019)
ISLR	Logistic regression software.	James et al. (2017)
pROC	ROC curve and AUC retrieval.	Robin et al. (2011)
iml	Derivation of the feature importance.	Molnar and Bischl (2018)
zoo	Performs last observation carried forward for missing data imputation.	Zeileis and Grothendieck (2005)

Appendix B: Respiration frequency domain analysis

Figure 1 shows the retrieval of respiratory rate from raw respiration data as explained in section 3.2.1. The plot represents one pilot during the SS experiment.

Figure 1: The retrieval of the respiratory rate from respiration data.



Appendix C: Task 1: Confusion matrices

In Table 1 the confusion matrix indicates the results of cognitive state prediction on the training and validation dataset. Table 2 indicates the confusion matrix for both pilots of crew 4 and 9. Table 3 and 4 indicate the results of the same flight crews for the benchmark model.

Table 1: Confusion matrix for the training and validation dataset predictions.

	Actual			
Prediction	BL	SS	CA	DA
BL	141963	2009	15244	6592
SS	526	4230	2	0
CA	380	34	56689	6
DA	758	0	1	5370

Table 2: Confusion matrix for predictions on the test dataset (crew 4 and 9).

	Actual			
Prediction	BL	SS	CA	DA
BL	580293	29035	279534	42494
SS	0	0	0	0
CA	65524	1705	88152	9950
DA	4920	0	0	1129

Table 3: Confusion matrix for the benchmark predictions on the training and validation dataset predictions.

	Actual			
Prediction	BL	SS	CA	DA
BL	123094	2790	7785	5942
SS	2329	3325	24	0
CA	10451	100	72209	1660
DA	225	26	0	3844

Table 4: Confusion matrix for the benchmark predictions on the test dataset (crew 4 and 9).

	Actual			
Prediction	BL	SS	CA	DA
BL	454657	21285	259702	38470
SS	3	0	0	0
CA	191965	9375	106884	14972
DA	2041	80	32	131

Appendix D: Task 1: Pilot dependent results

Figure 1: Cognitive state prediction for the pilot of crew 9, in the right seat.

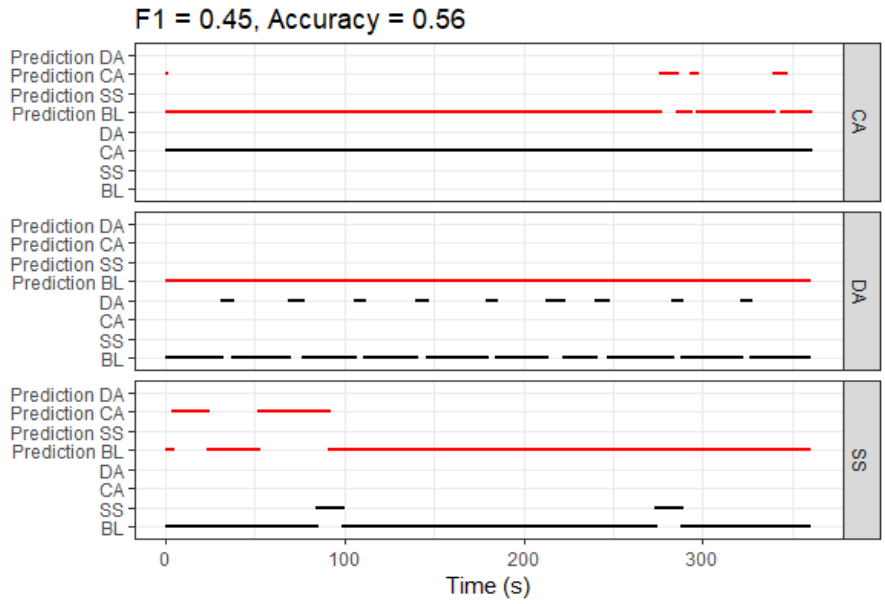


Figure 2: Cognitive state prediction for the pilot of crew 4, in the left seat.

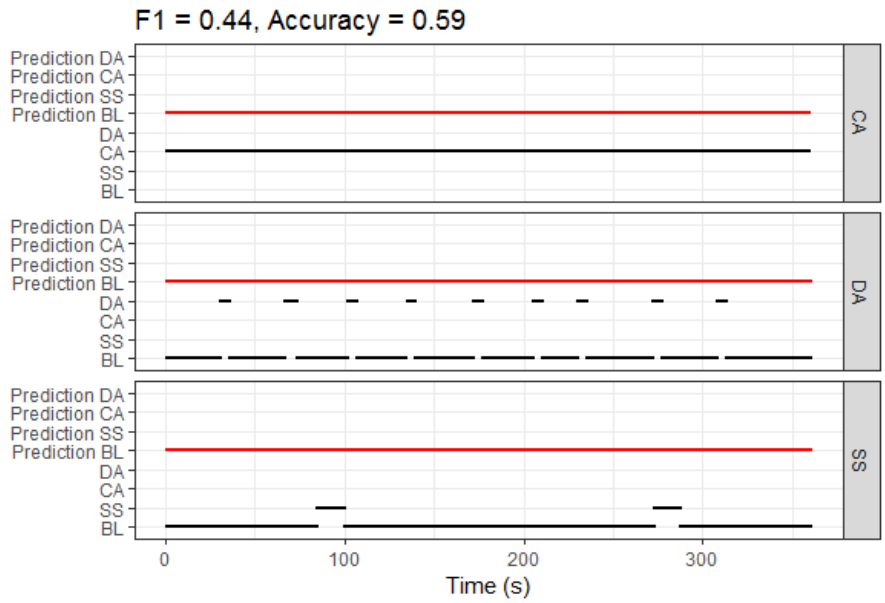


Figure 3: Cognitive state prediction for the pilot of crew 4, in the right seat.