DATA GENERATION FOR AI-BASED HUMAN PERFORMANCE EVALUATION FOR FLIGHT CREWS

P. Lorrig*, M. Ilic†, M. Biella‡, Z. Daw*
* University of Stuttgart, Institute of Aircraft Systems, Pfaffenwaldring 27, Stuttgart, Germany
† University of Brunswick, Institute of Flight Guidance, Hermann-Blenk-Str. 27, Brunswick, Germany
‡ German Aerospace Center, Institute of Flight Guidance, Lilienthalplatz 7, Brunswick, Germany

Abstract

AI-models could be leveraged for human performance evaluation with high accuracy, reliability and objectivity. Existential for properly training and applying AI-models are relevant, sufficient and reliable data. Two flight simulator studies were conducted with the aim to measure and collect flight but foremost physiological data from flight crews during flight deck activities. In the Air VEhicle Simulator (AVES, DLR Braunschweig) certified A320 flight crews ($n = 42$) performed various short flight scenarios, allowing a controlled experiment environment for data collection. These data were then post-processed in order to get a time synchronous signal between ECG and stress level feedback. Data from both studies were analyzed and a paired-sample t-test was conducted for the second study's data. Data collected during that study showed significant differences between the two Scenarios (Baseline and Stress) in feed-backed stress levels ($p < .001$) and measured heart rates ($p < .05$). Nevertheless, there are still recognizable gaps in duration and intensity of high stress as well as in gender diversity. Overall, about $39\,\mathrm{h}$ of physiological data have been recorded from flight crews during flight deck activities which are made publicly available.

Keywords

Human Performance; AI Systems; Human Factors

NOMENCLATURE

Indices

bl	Baseline Scenario
cpt	Captain
fo	First-Officer
hr	Heart Rate
l	Landing Section of the Stress Scenario
sc	Stress Scenario
sl	Stress Level

Abbreviations

AVES	All Vehicle Simulator
df	Degree of Freedom
ECG	Electrocardiography
EEG	Electroencephalography
LoHP	Limits of Human Performance
Mdn	Median Value
M	Mean Value
N	Number/Amount of
OSAT	Online Stress Assessment Tool
SD	Standard Deviation
SPO	Single Pilot Operation

1. BACKGROUND

Future aircraft systems, becoming continuously more automated, will require, as a basis for a human centred automation, the evaluation of flight crew performance capabilities. Adaptation of task-load according to the momentary performance capability of flight crews in such aircraft is just one possible use-case among many. Human performance evaluation systems are expected to detect a degradation in flight crew performance even before critical situations might occur. Hence, allowing timely interventions, better crew resource awareness, and enhancing overall flight crew performance as well as crew resource management (CRM) during flight. Therefore, the evaluation of human performance in real time with high accuracy becomes an urgent need for Human-AI-Teaming-Systems (HATS). Machine Learning models can be leveraged to perform this evaluation task with high accuracy, reliability and objectivity, yet their performance depends on relevant, sufficient and reliable data. This work focuses on the collection and quality assessment of such data to enable an AI-based approach for human performance evaluation in aviation.

2. CONCEPT OF STRESS DETECTION SYSTEM

Previous research has demonstrated a correlation between physiological signals and task-related workload levels, as evidenced by studies conducted by Riaz et al. [1] and Meteier et al. [2]. Furthermore, Taelman et al. [3] conducted a study involving 28 subjects, revealing an increase in heart rate during cognitive task-load increase. A variety of physiological signals are suitable for such analysis, including but not limited to Electrocardiogram (ECG) [4], Electroencephalogram (EEG), and functional near-infrared spectroscopy (fNIRS) [5].

doi: 10.25967/610521

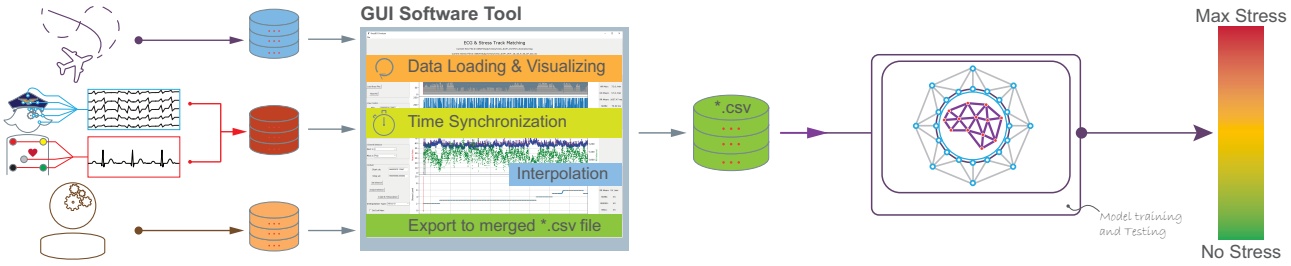

FIG 1. Conceptual structure of the process of data collection and processing applied in the second study (Limits of Human Performance) as well as the further usage.

PhysioNet[1], has two findings for ECG records under cognitive tasks. The data set by Amin et. al. [6] uses a wearable device to collect various data (heart rate, blood preassure, etc.) during exam situations. More interestingly, Healey et. al. [7] published a data set of ECG recordings of automobile drives in different street settings (e.g. city, highway, etc.). Although the data set by Healey et. al. looks promising the difference in training of the participant, tasks that had to be performed, and overall situation are quite different compared to the flight deck environment. Kaggle[2], has no findings for stress related physiological signals. Unfortunately, there is no data set available that consists of recorded physiological data of flight crews.
In this context, our focus during this study centered on the collection of ECG data and the self-reported stress levels of A320-type-rated pilots. Thus, a pre-trained machine learning model can be employed to analyze human performance based on these physiological signals, see Figure 1.

3. METHODS FOR DATA GENERATION

For generating physiological data from flight crews during flight deck activities two flight simulator studies were conducted in the AVES (see Fig. 3) at the German Aerospace Center in Brunswick, Germany. In order to stimulate stressful situations, we created unexpected deviations from normal flight pattern with various selected and triggered system failures (see Figure 2 and 3).

3.1. Single Pilot Operation Study (SPO)

A first study was performed using the data generated by Booms [8], which focuses on the behaviour of flight crews in single pilot operation during five different approach scenarios. For investigating this research question participants ($N_{SPO} = 24$), further demographics in Table 1, had to perform five different approach scenarios with different interferences and failures, see Figure 2.
In this study a 1-lead ECG (Medlab, EG01000) was recorded and each participant filled out a paper-based stress level feedback as well as a NASA TLX questionnaire after each flight.

3.2. Limits of Human Performance Study (LoHP)

Based on the outcomes of the SPO study requirements for a second study were derived in order to increase stress level feedback quality as well as induced stress during the flight scenarios. Furthermore, enabling a reliable time-synchronization between the different measured values especially stress levels and heart rates. Therefore, a new tool was developed, Online Stress Assessment Tool (OSAT),in order to request in a high frequency the momentary stress level of the participant, further explained in Section 3.2.1.
In this study, 9 flight crews ($N_{LoHP} = 18$) participated consisting each of a Captain ($N_{LoHP,cp} = 5$) and First Officer ($N_{LoHP,fo} = 13$), four crews consisted of two First Officers who formed a flight crew. Relevant demographic details are shown in Table 2.

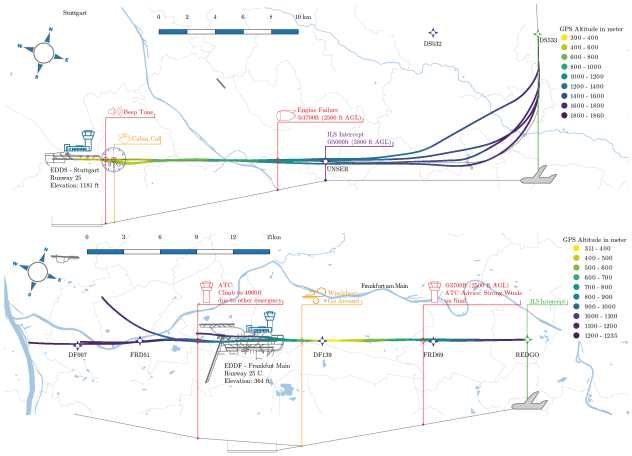

FIG 2. Example Scenarios of the SPO Study showing the tracks of different approaches by different participants.

	M	Mdn	SD
Age	32.1	31	5.28
Flight Hours			
- last 12 months	281.69	263	214.15
- total	2848.65	2450	2337.06
- on A320-type	2394.69	1600	2208.98

TAB 1. Demographic Data from SPO study

	M	Mdn	SD
Age	38.1	35.5	8.99
Flight Hours			
- last 12 months	378	475	251.88
- total	7363.05	4718	5653.70
- on A320-type	4313.89	2950	3480.34

TAB 2. Demographic Data from LoHP study

[1] PhysioNet is a website providing a wide range of physiological signal databases and related software tools for researchers and healthcare professionals to advance the field of biomedical data analysis and healthcare technology.

[2] Kaggle is an online platform that hosts data science competitions, provides data sets for machine learning practice, and offers a community for data scientists and AI enthusiasts to collaborate and learn.

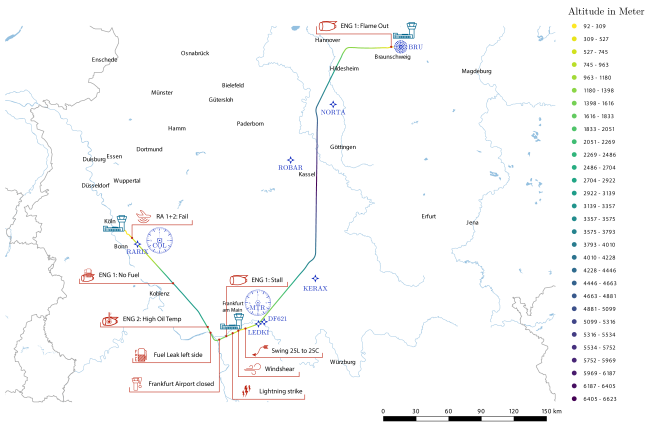

FIG 3. Schematic overview of the Stress Scenario of the LoHP Study with the different triggered failures and interferences.

During this study a 7-lead ECG with an additional respiration curve was recorded (Medlab, EG05000 + Respiration Board). Moreover, after briefing the flight crews for their flights a resting ECG was recorded for $5\,\mathrm{min}$ in order to get a physiological baseline. Subsequently, each flight crew performed two scenarios:

1) **Baseline Scenario:**
   From Brunswick Airport (EDVE) to Hamburg Airport (EDDH)
   Was used to record data from a flight with no failures or interferences.
2) **Stress Scenario:**
   From Brunswick Airport (EDVE) to Frankfurt Main Airport (EDDF)
   Consists of several interferences in order to trigger a high workload and thus stress, see Figure 3.

A single Stress Scenario was created based on the NASA TLX feedback given during the SPO study as well as the outcomes from the Future Sky Safety Project 6 Human Performance Envelope [9].
One scenario was deliberately selected instead of multiple scenarios for two primary reasons:

- To give flight crews an as close as possible flight experience and hence getting them mentally more into the simulation.
- To create an continuous scenario where complexity and task-load increases and hence a physiological stress reaction can build up.

3.2.1. Online Stress Assessment Tool (OSAT)

The OSAT was developed based on the Instantaneous Self-Assessment Scale (ISA) and the visual analogue scale (VAS) and was used for the first time in the LoHP study. This web-tool was then accessible during flight via the electronic flight bag (EFB) on the side of the flight deck positions, see Figure 5. Participants could either slide or tap on the corresponding stress level on an interactive slider and then needed to confirm the selected value with a button next to the slider, see Figure 4. The additional confirmation step - pressing the confirm button - was implemented to have a validation task in order to assess the actual workload.
The left update indicator turned red, increased in size, and started flashing after $1\,\mathrm{min}$ of no new confirmed value in order to remind the participant to give an update on their momentary stress level, see Figure 4 on the bottom. After another minute the slider started to slowly increase by 0.1 each

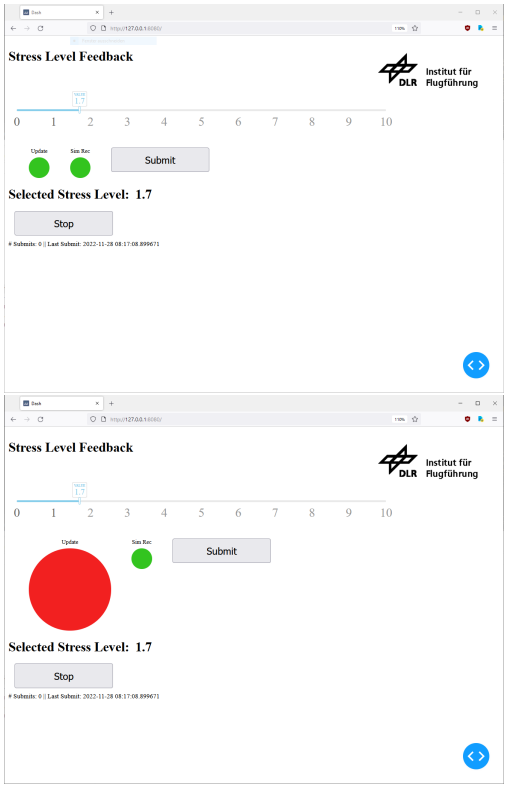

FIG 4. Online Stress Assessment Tool OSAT. top: Normal user interface during recording, the slider can be used to select the momentary stress level; bottom: User interface when no update has been given after $1\,\mathrm{min}$. The red indicator was flashing in the real installation to catch more attention.

FIG 5. Integration of the OSAT inside the AVES A320 flight deck. left: close-up of the captain-side tablet.

$2\,\mathrm{s}$. This indicator increase aims to emulate an increase in workload as tasks get neglected and hence, the momentary stress level must be have increased to the previously selected one as well.

4. EXPLORATORY DATA ANALYSIS

In order to comprehend the composition and distribution of the acquired data an exploratory data analysis (EDA) was performed. In three major steps different aspects of the stress level and hear rate data were analysed and compared between the Baseline and Stress Scenario.

1) Histogram plots are used to analyse the overall distribution.
2) Heat map plots with normalized run-times are used for time relative analysis.
3) Paired-sample t-tests are conducted to analyse the statistical impact.

The first aspect to investigate was the general distribution of the input data (ECG-Signal) and reference labels (stress

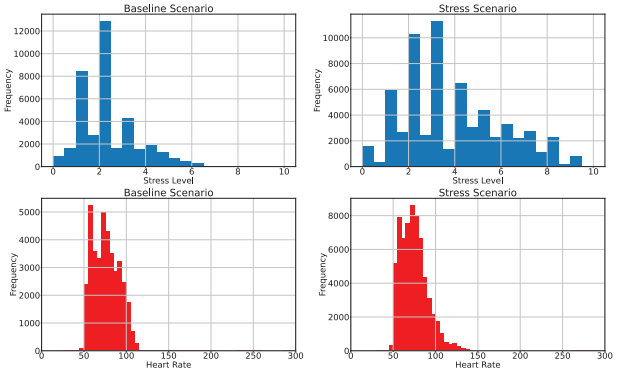

FIG 6. Histograms showing the overall distribution of the stress level (top) and heart rates (bottom) for the Baseline (left) and Stress Scenario (right).

level). The feedbacked stress levels and momentary heart rates were then plotted in a histogram each, see Figure 6.
The next aspect to analyse was how input data and reference labels are distributed over the run-time. As the run-time was different for each record, each signal was normalized in run-time to be between 0 and 1 ($t_{norm} \in (0, 1)$). Only a minor error is introduced by this, given that the relative flight progression is consistent throughout every crew, see Figure 7. This consistency stems from the fact that all flights started and ended at the same airport and experienced identical interference events at the same points in time over the flight trajectory.
The heat map plot in Figure 8 and 9 shows the stress levels and heart rates against the normalized run-time, respectively. This means that each participants data was interpolated to the same number of discrete samples. The brighter a section is the more frequent a certain value was present at that time over all participants records. Additionally, the mean value of each signal over all participants was calculated and plotted as well.
In order to validate statistical differences between the Baseline and Stress Scenario a paired-sample t-Test was performed with the following hypothesis:

- **Null Hypothesis (H0):** There is no significant difference in the stress levels between the Baseline and Stress Scenarios after manipulating the workload.
- **Alternative Hypothesis (Ha):** There is a significant difference in the stress levels between the Baseline and Stress Scenarios after manipulating the workload.

Both, stress levels and heart rates have then been resampled to have the same sample size by applying a 1-D linear interpolation. These were then used as input values for a paired-sample t-Test.

5. RESULTS

5.1. SPO

Since the stress level feedback was done after each flight on a very coarse track no reliable time-synchronization between the stress level and ECG recording could be established. This made it impossible during post-processing to reliably match the two signals. The following requirements have emerged from this study:

1) A reliable time synchronization between the stress level feedback and physiological signal must be established.

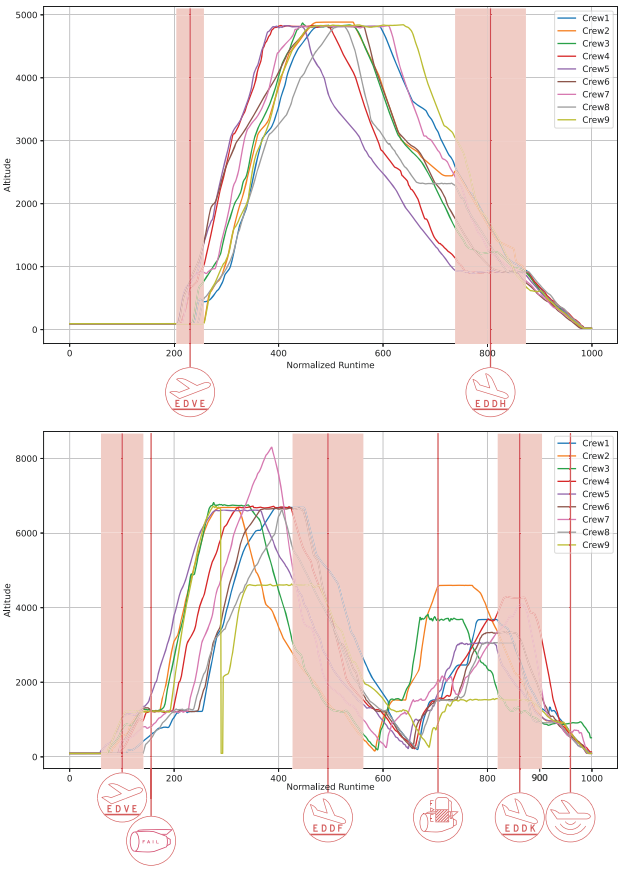

FIG 7. Top: Baseline; Bottom: Stress Scenario. Simulator flight data showing the GPS Altitude of all flight crews. The area from first to last is indicated with the light transparent red box. For Crew 9 during the Stress Scenario the simulator had to be restarted resulting in this ditch.

2) Improvement of quality and frequency of the stress level feedback. This means more often and correlated to the exact point of time.
3) Overall perceived stress needs to be increased. This is done by introducing one flight as well as various interference actions and degradation of aircraft automation.
4) Flight scenarios need to be as realistic as possible in an accurately behaving aircraft simulator.

5.2. LoHP

The paired-sample t-Test showed that during the Stress Scenario ($M_{sc,sl} = 3.7$, $SD_{sc,sl} = 0.18$) and Baseline Scenario ($M_{bl,sl} = 2.3$, $SD_{bl,sl} = 0.075$) each participant had a significant higher stress level ($t_{mean}(45) = -9.488$, $p_{mean} = 1.911 \times 10^{-5}$, $df = 99$). The effect size according to Cohen (1992) [10] is on average $r_{mean} = .64$ and represents a strong effect. This indicates that the visible assumption that more stress had been perceived by participants compared between Baseline and Stress Scenario can be statistically demonstrated.
Subsequently, we explored whether there are any variations within the physiological responses using momentary heart rates. The entire run-time was analyzed first revealing that the mean heart rates during the Baseline Scenario were higher than during the Stress Scenario. The paired-sample t-Test shows that within the Stress Scenario ($M_{sc,hr} = 72.41$, $SD_{sc,hr} = 7.25$) and Baseline Scenario ($M_{bl,hr} = 73.42$, $SD_{bl,hr} = 7.02$) the heart rate was on average higher ($t_{mean}(45) = 3.68$, $p_{mean} = .0227$,

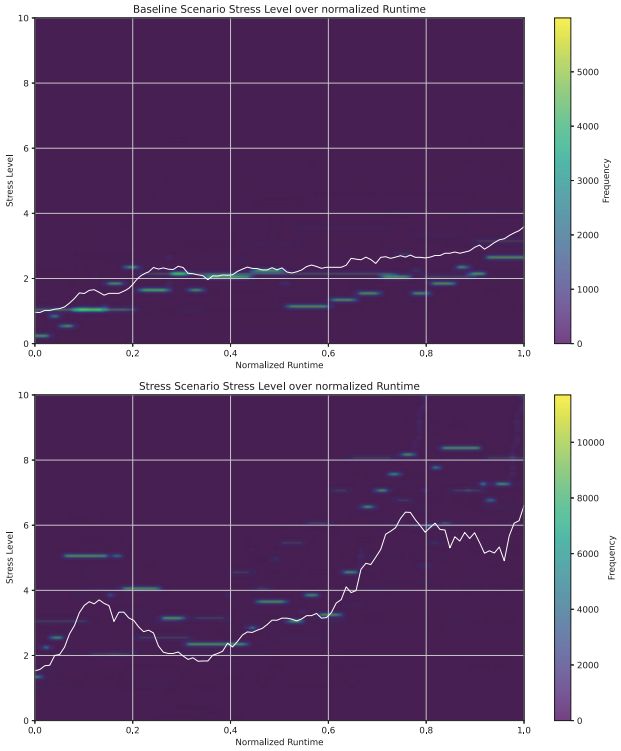

FIG 8. Heat maps of the distribution of stress levels over the normalized run-time.

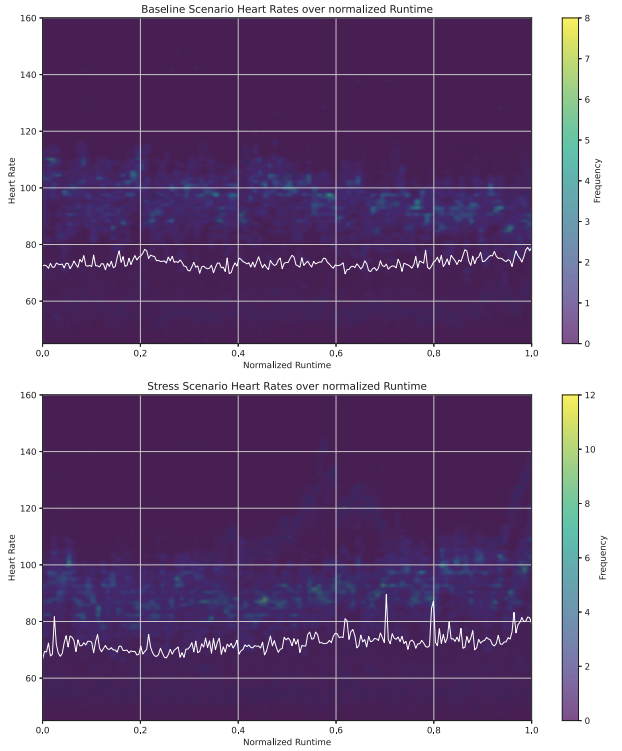

FIG 9. Heat maps of the distribution of heart rates over the normalized run-time.

$df = 999$) with an medium to low effect size of an average of $r_{mean} = .25$.

Although this finding might be contra intuitive, this results are consistent with considerable amount of uneventful flight during the stress scenario, namely the flight time from Brunswick up to the approach to Frankfurt Airport. Therefore, the recordings were analyzed with the same methods but close to the end of run-time where the landing section is to be assumed ($t_{norm,Land} \in (0.95, 1)$). These values were chosen by plotting the altitude of all flight simulator data. Furthermore, it is assumed that also the most stress has build up within the Stress Scenario since up to this point as many flight systems have failed or become unavailable.

For the landing section the paired-sample t-Test shows that during the Stress Scenario ($M_{sc,hr,l} = 78.40$, $SD_{sc,hr,l} = 6.38$) and Baseline Scenario ($M_{bl,hr,l} = 75.43$, $SD_{bl,hr,l} = 5.27$) the heart rate was on average higher ($t_{mean,l}(45) = -2.73$, $p_{mean,l} = .0175$, $df_l = 49$) with an high effect size of an average of $r_{mean,l} = .66$. However, five Participants had on average a significant higher heart rates during the Baseline Scenario still. This might be due to a change in commands from pilot flying to pilot monitoring[3] over both scenarios. An explanation for the difference compared to the stress level might be that those pilots have, even as pilot monitoring, noticed an increase in workload which was feed-backed through the OSAT, but did not, at least not to a measurable degree, where it would induce an increase in heart rate.

It is important to highlight that this assumption is grounded in the concept of a linear correlation between an elevation of heart rates and the presence of stress as demonstrated by Shubert et. al. [11].

Anyhow, both of these paired-sample t-Tests statistically verify that there is a significant difference between both scenario runs. For further and deeper analysis, it might be also important to consider which flight crew member shared what tasks as this might impose another level of workload for each member.

[3] Pilot flying (PF) is responsible for physically operating the aircraft, while pilot monitoring supports PF by managing communications, systems, and monitoring flight parameters during a flight.

6. DISCUSSION

There are still persistent gaps that need to be addressed and mitigated before considering training a commercial machine learning network. First and most importantly, during the second study no female person was found to participate. Leaving a huge gender data gap that might be especially significant as the physiology between genders also varies. This means that when using such systems with a bias towards male data the performance might be significantly worse or might even give out false classifications when using on female persons.

Another data gap which might not be as severe as the other is within the physiological stress reaction. Although, it is possible to find statistical significant differences between Baseline and Stress scenario, there is still the need for more and higher stress that persists for a longer period of time. Especially the stress needs to be to a degree where significant physiological reactions occur. A stress detection system should be capable of detecting critical situations before they have a negative effect on flight safety. Therefore, it is important to have physiological data not only under high stress up to a potential loss of control but also before that and how it build up.

It is important to emphasize that, during the Stress Scenario, despite the aircraft being in a very degraded state on final approach towards Cologne Bonn Airport, all flight crews displayed exceptional competence in managing the situation and successfully executed a safe landing and touch-down. This indicates that the flight crews were trained extremely well allowing them to achieve landing under such bad conditions. This leads to another aspect of improvement: the

flight scenario needs to be modified so that it becomes more challenging but at the same time not becoming unrealistic or leaving too few options for flight crews.
For training a machine learning network, a more homogeneous data distribution can mitigate model training biases. This means that all labels should be present in almost the same amount for a specific training data set. With the current data set this can only be achieved by cutting out certain areas (e.g. segments with no interference). On the other side very high stress levels were underrepresented. The design of the scenarios is crucial for the success of such a data recording.

7. CONCLUSIONS

Over the course of both studies 39 h17minutes of physiological data have been collected (Single Pilot Operation study: 10 h 58 min; Limits of Human Performance study: 28 h 19 min).
To the best of our knowledge, this is the first set of physiological data from flight crews under several conditions have been recorded, analyzed and evaluated.
The data collected from both studies as well as software created and used for analysis can be found in a GitHub repository[4] for further open usage.
The Exploratory Data Analysis shows a significant difference between the Baseline and Stress Scenario of the Limits of Human Performance study. This suggests that participants not only experienced higher level of stress during the Stress Scenario but also to a quantifiable measure, as evidenced by the recorded ECG data.
Future work can further apply signal analysis by other established parameters and methods e.g. HRV-Analysis[5]. Moreover, the need to mitigate the persisting gender data and high stress level gaps are aspects to consider in future simulator studies for data collection.
To ensure a robust machine learning driven system for human performance evaluation with high efficiency and accuracy, the identified gaps need to be addressed either by data augmentation or further studies to gather data.

Contact address:

patrick.lorrig@ils.uni-stuttgart.de

References

[1] Roha Riaz, Nuzhat Naz, Mnahil Javed, Farhat naz, and Hamza Toor. Effect of mental workload related stress on physiological signals. In *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. IEEE, mar 2021. DOI: 10.1109/iecbes48179.2021.9398833.

[2] Quentin Meteier, Emmanuel De Salis, Marine Capallera, Marino Widmer, Leonardo Angelini, Omar Abou Khaled, Andreas Sonderegger, and Elena Mugellini. Relevant physiological indicators for assessing workload in conditionally automated driving, through three-class classification and regression. *Frontiers in Computer Science*, 3, jan 2022. DOI: 10.3389/fcomp.2021.775282.

[3] J. Taelman, S. Vandeput, A. Spaepenand, and S. Van Huffe. Influence of mental stress on heart rate and heart rate variability. *4th European Conference of the International Federation for Medical and Biological Engineering*, 2008.

[4] J.A. Healey and R.W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):156–166, jun 2005. DOI: 10.1109/tits.2005.848368.

[5] Anneke Hamann and Nils Carstengerdes. Assessing the development of mental fatigue during simulated flights with concurrent EEG-fNIRS measurement. *Scientific Reports*, 13(1), mar 2023. DOI: 10.1038/s41598-023-31264-w.

[6] Md Rafiul Amin, Dilranjan Wickramasuriya, and Rose T Faghih. A wearable exam stress dataset for predicting cognitive performance in real-world settings, 2022. DOI: 10.13026/KVKB-AJ90, https://physionet.org/content/wearable-exam-stress/1.0.0/.

[7] Jennifer A Healey and Rosalind W Picard. Stress recognition in automobile drivers, 2008. DOI: 10.13026/C2SG6B, https://physionet.org/content/drivedb/.

[8] Christian Booms. Consequences of an increasing workload for single pilots during (non-) precision instrument approaches -a flight simulator study. Master's thesis, Universität Ulm, 2023. https://elib.dlr.de/194082/.

[9] Sara Silvagni. Concept for human performance envelope. Technical report, Future Sky Safety, 2015.

[10] Jacob Cohen. Statistical power analysis. *Current Directions in Psychological Science*, 1(3):98–101, June 1992. DOI: 10.1111/1467-8721.ep10768783.

[11] C. Schubert, M. Lambertz, R.A. Nelesen, W. Bardwell, J.-B. Choi, and J.E. Dimsdale. Effects of stress on heart rate complexity—a comparison between short-term and chronic stress. *Biological Psychology*, 80(3):325–332, Mar. 2009. DOI: 10.1016/j.biopsycho.2008.11.005.

[4] https://github.com/0Patrice/StressDetectionInFlightCrews

[5] Heart Rate Variability (HRV) is a measure of the variation in time intervals between successive heartbeats, used to assess the autonomic nervous system's activity and overall physiological health.