





# Moving Beyond Physiological Baselines: A New Method for Live Mental Workload Estimation

Torsten Gfesser<sup>(✉)</sup> , Thomas E. F. Witte , and Jessica Schwarz 

Fraunhofer FKIE, Fraunhoferstr. 20, 53343 Wachtberg, Germany  
{torsten.gfesser, thomas.witte,  
jessica.schwarz}@fkie.fraunhofer.de

**Abstract.** The analysis of physiological data can provide valuable information on the mental state of users interacting with a technical system, such as an intelligent tutoring system. By obtaining live estimations of mental workload a learning system can adapt, e.g., the level of difficulty of tasks to the learners needs. However, the analysis and interpretation of physiological data usually requires a baseline recording at a rested state prior to or after a task limiting their practical value. Additionally, the baseline of a physiological measure cannot be considered as a stable value but varies between days and even within a day interpersonally, so the validly calibrated data of a baseline become invalid over time limiting its value for long term use cases.

This paper proposes a new method for near real time mental workload estimation. A machine learning model which predicts the mental workload based on the heart rate variability (HRV) derives metrics without the necessity of baseline recordings. First, a machine learning model is trained on a dataset of previously collected physiological data and corresponding mental workload ratings. Subsequently, physiological measures are collected continuously from a participant throughout tasks. The model is then used to predict the participant's mental workload in real time based on the HRV data.

The results of our pilot study show first empirical support, that the proposed analysis technique is able to estimate mental workload in near real time with an accuracy of 90%.

As this technique does not depend on baseline recordings it has the potential to be specifically valuable in applied settings such as adaptive training systems or to monitor the mental health of workers in safety-critical industries. The method could also be extrapolated for the analysis of other physiological measures in future research.

**Keywords:** Mental Workload · Baseline · Real Time · Physiological · HRV · Artificial Intelligence

## 1 Introduction

Live estimation of mental states, such as mental workload, is an important requirement for the design of adaptive technical systems that adapt their behavior to the current state of the user. As an example, adaptive systems may use live detection of mental

workload to support the operator, if a critically high level of mental workload, has been detected. In the learning context such information could be used by an adaptive training system to specifically provoke states of high workload to train the learner how to cope with critical conditions. Measuring mental workload by physiological measures such as heart rate, heart rate variability, or pupil size is one of the most prominent approaches in this context. Benefits compared to e.g., subjective ratings are that most of these measures can be recorded continuously during a task without disturbing the user and data can be analyzed in near real time. However, physiological reactions can differ strongly between and even within individuals depending on the fitness level, caffeine consumption, and physical activity among others. A common approach to account for inter- and intraindividual differences is recording a baseline at the beginning or at the end of a task in a relaxed state and comparing recorded data during a task with the baseline value. However, in real-world applications baseline recordings are not always a suitable method as there is often not enough time for a baseline recording. Also, if baselines must be recorded regularly to be able to work with a technical system, this can be disturbing and often lowers the user acceptance of the system.

In this paper we introduce a new method for live mental workload estimation without baseline recordings, making it more applicable for real-world settings. We used the heart rate variability (HRV) as a physiological measure to develop this method, because it is considered as an established indicator of mental workload and stress.

Section 2 gives an overview on prior studies on HRV assessment. Section 3 introduces our new method for live analysis, also describing the data used for validation. Section 4 describes the AI Classifier, which was trained for the live analysis of mentally demanding tasks based on the calculated metrics from Sect. 3. This paper ends with a discussion, the limitations of the method as well as conclusions and future developments.

## 2 Heart Rate Variability as an Indicator of Mental Workload

Mental workload refers to the amount of mental effort and resources required to perform a specific task, encompassing cognitive processing, attention, and effort exerted by an individual while engaging in a task [1]. It is a measure of the cognitive and perceptual demands of a task, influenced by factors such as task complexity, time pressure, and environmental conditions [2]. Heart rate variability (HRV) analysis measuring the variation of time intervals between heartbeats is a valuable method for assessing autonomic function of the cardiovascular system. As such it is often used as a potential biomarker for various health conditions [3] and for predicting mental workload in various settings.

The HRV can be calculated using various methods, including time-domain and frequency-domain. In time-domain analysis, parameters such as SDNN (standard deviation of NN intervals) and RMSSD (root mean square of the differences between adjacent NN intervals) are commonly used [4]. Frequency-domain methods involve the use of spectral analysis, such as fast Fourier transform and autoregressive model, to calculate parameters like high frequency and low frequency components [5]. The Task Force of the European Society of Cardiology recommends the use of SDNN and RMSSD as widely adopted measures of HRV [4].

Literature shows that there is a negative correlation between the SDNN and subjective mental workload [6], which also applies to the RMSSD [7]. Delliaux et al. [8]

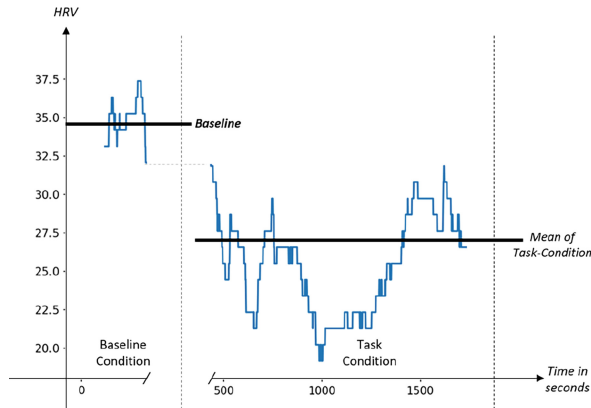
and Radüntz et al. [9] have focused on characterizing the impact of mental workload on cardiovascular function using HRV non-linear indexes and inherent timescales of cardiovascular biomarkers, providing insights into the relationship between HRV and mental workload. Their results indicate that mental workload significantly lowered the non-linear dynamics of RR interval [8] and that the assessment of mental workload using cardiovascular biomarkers' inherent timescales provide valuable insights into the physiological responses associated with varying levels of cognitive demand [9]. Also, Forte et al. [10] concluded in their systematic review about the relationship between HRV and cognitive functions, that HRV is closely linked to cognitive function. Veltman and Jansen [11] emphasized the differentiation of mental effort measures and its consequences for adaptive automation, highlighting the importance of HRV in assessing cognitive workload. The physiological basis of HRV as a reflection of autonomic nervous system activity and its role in emotion regulation further supports its relevance in adaptive systems, as highlighted by Witte et al. [12].

Veltman and Gaillard [13] concluded that HRV is a sensitive index for mental workload when tasks are highly demanding, emphasizing its relevance in assessing cognitive demand during complex tasks. Research by Cinaz et al. [14] and Shao et al. [15] has focused on the use of HRV in monitoring mental workload levels during office-work scenarios and human-robot interaction, respectively, highlighting the versatility of HRV in diverse domains and environments. Shao et al. [15] conducted a comparison analysis of different time-scale HRV signals, demonstrating the applicability of HRV in evaluating cognitive demand during interactive tasks. This supports the notion that HRV can discern fluctuating task demands and attenuate during mentally straining workloads, as stated by Nardolillo et al. [16].

The findings from these studies collectively highlight the potential of HRV as a valuable physiological marker for assessing mental workload during software tasks, providing insights into cognitive demand and adaptive responses in various task environments.

## 2.1 Physiological Baselines

One of the most common methods to account for intra- and interindividual differences in physiological measures, such as HRV, is to use baselines. This method is considered as the gold standard for analyzing HRV and other physiological measures, where the “tonic level measured immediately prior to stimulation is referred to as the baseline, the level of activity against which we compare the phasic response to a stimulus” [17]. A baseline can be obtained by taking the average of HRV values over a specified time interval, such as five minutes, under resting or non-stressful conditions before or after a task [18]. This baseline can then be used to compare it against HRV values collected during physical or mental tasks. The method of obtaining a baseline is visualized in Fig. 1.



**Fig. 1.** Baseline for HRV obtained in the first five minutes (300 s) in comparison to the HRV mean of the task condition.

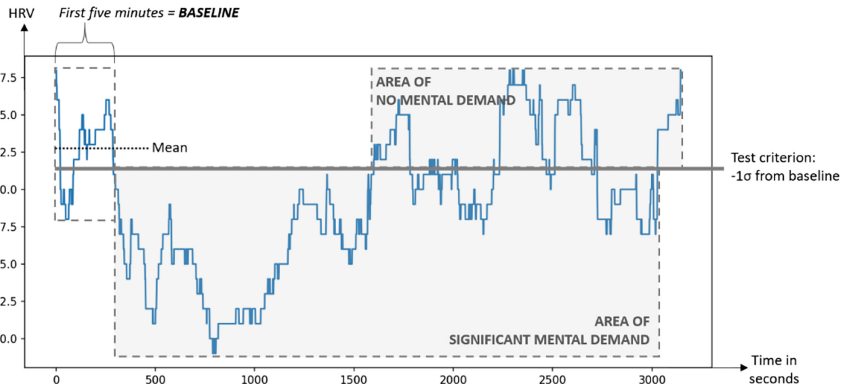
## 2.2 Challenges of Live HRV Analysis

While HRV is a valuable health metric, it has certain peculiarities that make it difficult to establish a definitive baseline that is valid over long periods. When it comes to baseline-measurements the respective physiological variable should be stable across the analyzed period. Regarding short-term measurements, Tarkiainen et al. [19] conducted a study on the stability over time of short-term HRV, indicating that most short-term HRV measures were highly stable over time in laboratory conditions. They further conclude, that their SDNN obtained during 40-min recordings was more stable than the SDNN obtained during 5-min periods and the SDNN showed large variability in consecutive recordings. The stability and variance of HRV appear to vary depending on the period of recording.

Baseline measurements compare a short period of HRV values to another period of HRV values, similar to a consecutive recording. However, baseline measurements are implying that the physiological variable is stable in variance across time, regardless of the period. This assumption is important because it enables the comparison of the participant's response to a stimulus and their initially recorded baseline level. Physiological measurements can result in chaotic timeseries [20] violating this assumption. Chaotic timeseries refers to a type of time series data that exhibits chaotic behavior, characterized by instability, unpredictability, and sensitivity to initial conditions [21]. Without stability, there may be no anchor point in the psychophysiological data that could serve as a good baseline.

HRV baselines can become less valid over time because the range of values being evaluated can change. This may be the case when comparing consecutive recordings of short-term HRV. Individual HRV values vary greatly. Age, gender, health status, and even psychotropic substances like caffeine intake can influence the HRV. Secondly, intrapersonal homeostasis is a factor of influence, causing natural fluctuations even without external stimuli [22]. That means that a fixed baseline can shift throughout the day based on the current physical state of the individual. Comparing absolute values, like the baseline, across individuals or even comparing someone's current HRV to their own baseline, or any past measurements, can therefore be misleading. If the HRV changes

too much without external factors, there is a high chance of false positives (false alarms e.g., false classification of tasks as mentally demanding) or false negatives (misses, missed detections of mentally demanding tasks). This case is visualized in Fig. 2 where according to the initial baseline, huge parts of the following values would be classified as significant mentally demanding. A valid and reliable baseline must thereby shift based on the current physiological change, that is not related to external factors. It has to migrate with it as a measure, like normalizations do. A change in the current physiological state must be recognized and the shift of the baseline and a shift due to physiological or cognitive impact factors has to be differentiated.



**Fig. 2.** HRV values from one participant over a period of around 55 min, where the first five minutes are serving as the baseline. The test criterion by which HRV is classified as mentally demanding is HRV values below the first standard deviation ( $-1\sigma$ ) based on the baseline values.

The condition of the participant is usually unknown when the baseline is taken. If the subject is too excited at the beginning, then the HRV may only increase over time. If the subject is too relaxed at the beginning, then the HRV may initially lose altitude rapidly. This is called the law of initial value, where the magnitude of a response to a stimulus depends on the starting level of the measured variable [23]. For example, if a participant already has a very low HRV, which would also be measured as a baseline, then HRV-lowering stimuli may no longer be noticeable. However, if the participant has a very high HRV at the beginning, it may be that the entire test period is significantly below the initial value and therefore the effects of individual stimuli are no longer recognized.

Small or finer oscillations in the data are completely ignored when physiological baselines only compare the mean of a few conditions. Oscillations within a condition, based on single stimuli, will vanish through the calculation of means for whole experimental phases.

HRV is often used in retrospective analysis, where at least the values of the recorded timespan are completely available for state-of-the-art time series analysis techniques, such as removing possible trends or seasonality from the data. But without knowing the future data, an online or real time method, can only rely on the current data and data recorded in the past. Future data can't be forecasted because of the chaotic nature of

the HRV time series, as stated previously in this section. So, all useful metrics must be calculated based on the latest and passed HRV values.

The stated challenges are mostly valid for detailed live analysis with only HRV as a parameter. Therefore, several approaches like the multifactorial RASMUS Framework from Schwarz and Fuchs [24], use a combination of multiple parameters to provide more robust user state analyses.

Hoover et al. [25] tried to detect changes in mental workload based on real-time monitoring of HRV. Their original intent was to determine whether a change in task caused a change in HRV measurement. Using sub-Gaussian functions, they were able to successfully detect change points based on a change in tasks. This provided insight that mentally demanding tasks can be identified by changes in HRV. Further, they conclude that their method can successfully detect changes that are quite subtle.

However, to the best of our knowledge literature does not suggest a profound method for the live classification of short or ultra-short fluctuations of a single physiological parameter, like the HRV.

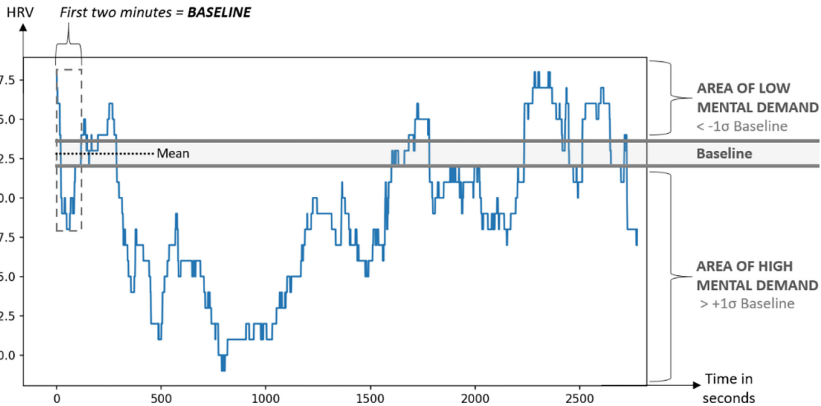
### 3 Introducing a New Method for Live Mental Workload Estimation

In this section we present a new method to enable a more detailed analysis of HRV fluctuations that may be useful e.g., for an efficient live analysis of workload in adaptive system design or for the evaluation of applications and tools in and outside of a laboratory. Our method aims at eliminating the need for regular baselines and providing a metric that is stable over time periods for classifying mental demanding fluctuations in the HRV.

#### 3.1 Re-analysis of Data from a Prior Study

A prior study conducted by Bruder and Schwarz [26] was used to develop the method of this paper by re-analyzing the data, where the HRV was calculated as a rolling 300 Heartbeat SDNN. A HRV baseline was calculated for each participant based on the first 120 s. During that time, the mental workload was kept low to moderate. This baseline was further used as a test criterion where a HRV value lower than one standard deviation from the baseline distribution will be classified as critically high and a value greater than one standard deviation as critically low mental workload when coinciding with a performance decrement. Figure 3 visualizes the concept and classification based on the data of one participant from the original study.

The initial study of the authors was followed by a validation study which confirmed that their used method is temporally valid and, moreover, could distinguish between three different conditions named baseline, high workload, and monotony. Workload assessment was based in this study on a combination of HRV and four other workload indicators (the number of tasks, number of mouse clicks, pupil diameter, and respiration rate) to compensate for inaccurate classifications of single indicators. The method proposed in this paper aims at developing this approach further by providing HRV-based live classifications of even short-term mentally demanding tasks, thus increasing the accuracy of this classifier.



**Fig. 3.** The classification of low, none and high mental demand based on a two-minute baseline, where a HRV value lower than  $1\sigma$  will be classified as critically high and a value greater than  $1\sigma$  as critically low mental workload.

### 3.2 Experimental Setting

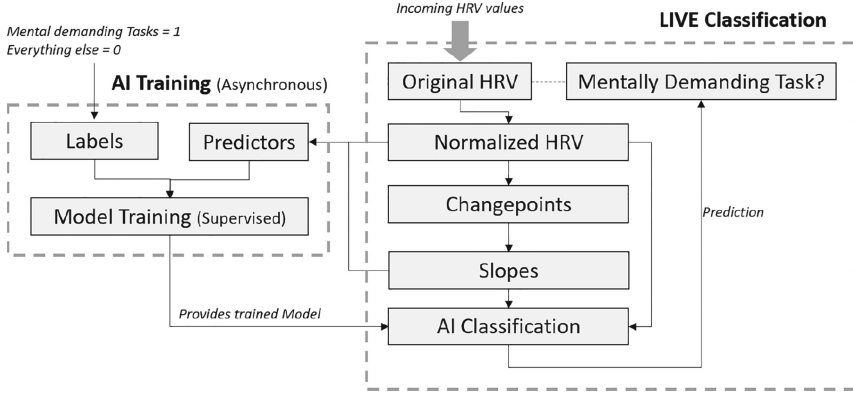
For the evaluation of the method provided in this paper, we utilized the existing data from the previous study conducted by Bruder and Schwarz [26] about the evaluation of diagnostic rules for real time assessment of mental workload within a dynamic adaptation framework. The framework was operationalized for an air traffic surveillance task.

The original study involved a sample of 15 participants (8 males, 7 female) aged between 20 and 51 years ( $M = 31.26 \pm 8.27$ ). A multisensory chest strap (Zephyr BioHarness 3) was used to collect data on HRV and respiration rate. Pupil diameter was recorded with an eye tracker placed underneath the monitor.

Participants began with a ten-minute training session where the examiner clarified task completion for each subtask. Following this, they engaged in a 45-min experimental test divided into three continuous phases, punctuated with a survey when a performance decrement on one of the tasks had been detected. In this survey, participants rated their perceived mental workload and other mental states. The experiment's duration therefore varied based on user performance, with additional workload ratings recorded both after training and at the experiment's conclusion to establish individual baselines [26].

### 3.3 Concept for a Live Analysis of Cognitive Workload

Steps for live analysis of the HRV data are normalizing the incoming HRV values and calculating additional metrics such as the slope of the ascending and descending HRV. Finally, the normalized HRV and metrics are used to be classified using a pre-trained machine learning model. The whole data flow is shown in Fig. 4. The steps specifically of the live classification will be explained in detail in the following section.



**Fig. 4.** The processing of incoming HRV values towards the prediction of mentally demanding tasks.

### 3.4 Live Normalization for Non-stationary Timeseries

It is important to normalize the HRV for comparability even of the same subject, because of the unpredictability of the future HRV values and regarding to the intraday differences.

In our method, we estimate a rolling normal distribution with a maximum likelihood estimation (MLE) based on the last 60 HRV values. The calculated mean from the distribution serves as the current plateau of the HRV, from which a new gradient will relatively be measured. Mathematically, the distribution means from the last 60 HRV values will be subtracted from the incoming HRV values, to get the relative value compared to the last minute in our case.

$$HRV = \{hrv_1, hrv_2, \dots, hrv_n\}$$

$$Rolling\ Mean = \frac{1}{60} \sum_{i=n-59}^n HRV_i$$

$$HRV_{Normalized} = \{hrv_1 - Rolling\ Mean, hrv_2 - Rolling\ Mean, \dots, hrv_n - Rolling\ Mean\}$$

This normalization forces the values to move around a center of 0 for each subject, which acts detrending, so it removes trends in the time series. Figure 5 plots the HRV values from one subject over a period of around 53 min with the normalized HRV values below.

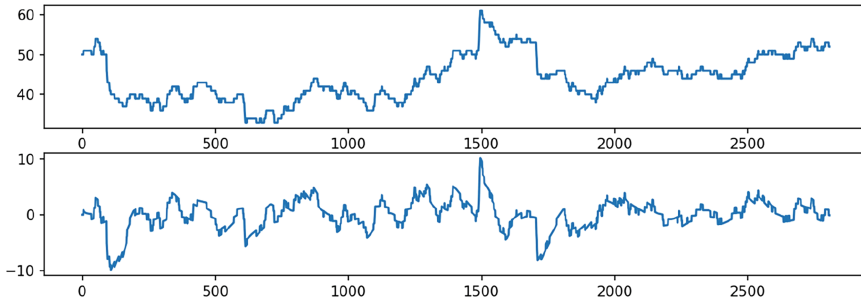
### 3.5 Online Change Point Detection

The second step is to recognize a change in the incoming flow of HRV values. Whenever there is a significant change, we must mark the position of that change point.

A comparison of two single values, so the last past value and the newest, won't be accurate in case of slightly fluctuating values. To address this problem, we always take the last twenty HRV values, where the first ten values are compared with a Wilcoxon rank sum test to the latest ten values, formulated as:

$$HRV_{Normalized} = \{hrv_1, hrv_2, \dots, hrv_n\}$$





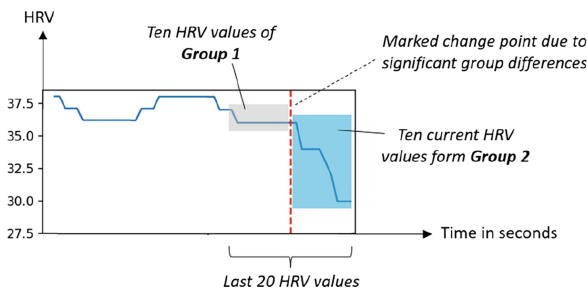
**Fig. 5.** Original HRV values and the normalized HRV values below.

$$HRV_{Group1} = \{hrv_{n-20}, hrv_{n-19}, \dots, hrv_{n-10}\}$$

$$HRV_{Group2} = \{hrv_{n-9}, hrv_{n-8}, \dots, hrv_n\}$$

$$W = WilcoxonRankSum(HRV_{Group1}, HRV_{Group2})$$

This is visualized in Fig. 6 are grouped colored in blue. In most cases, one or both groups do not meet the test assumptions of parametric inference statistical tests, such as the assumption of normal distribution. Therefore, the two groups are compared using a non-parametric Wilcoxon rank sum test with a fixed 5% alpha significance level.



**Fig. 6.** Calculation of change points based on the previous 1–10 values (Group 1) and the newest 11–20 values (Group 2).

In our method, the test criterion of the rank sum test will always be tested twice as one-sided tests for lower and greater significant difference. If the latest ten values are significantly greater, than an ascending changepoint will be reported. If significantly lower, a descending changepoint will be reported.

$$W_{Lower} = WilcoxonRankSum(HRV_{Group1}, HRV_{Group2})$$

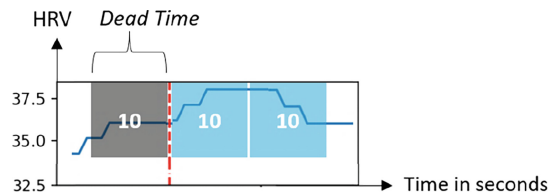
$$Descending\ Changepoint = \begin{cases} True & \text{if } W_{Lower} \text{ is } < 0.05 \alpha \\ False & \text{otherwise} \end{cases}$$

$$W_{Greater} = WilcoxonRankSum(HRV_{Group1}, HRV_{Group2})$$

$$Ascending\ Change\ point = \begin{cases} True & \text{if } W_{Greater} \text{ is } < 0.05 \alpha \\ False & \text{otherwise} \end{cases}$$

If there are two or more change points, each following the same direction, e.g., all lower or all higher, the series of all such change points is cached until the series is broken by a new change point with a different direction. The cached series then forms a single ascent or descent to be able to calculate a slope over the entire ascent or descent series.

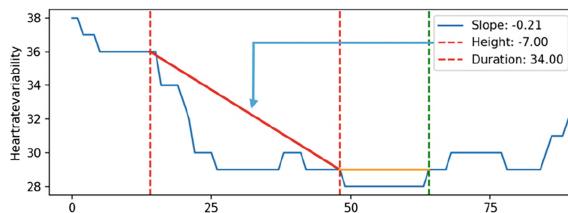
After a change point has been detected, a dead time of 10 HRV values starts in order to avoid that the already significant range of values is compared again. Figure 7 visualizes the dead time, where the first group of ten values was significantly different compared to the second group of ten values. So, a new change point was identified between the first and second groups. To avoid a new comparison of the first group, no group comparison is carried out as long as ten new values are available. After that, for each new value, the groups are calculated as specified in the paragraphs above.



**Fig. 7.** Dead time in the change point detection.

### 3.6 Calculating the Slopes

The third step is to calculate the slopes between the change points. Every time a new change point is detected, we can calculate the slope between the previous change point and the new one, as well as the slope over an entire series of ascent or descent change points. Figure 8 visualizes the calculated slope between two descending change points.



**Fig. 8.** Two visualized slopes between three change points. Red dotted lines are descending change points, green dotted lines are ascending change points. The red line is the ascending slope between two change points, followed by a yellow line indicating no slope. (Color figure online)

### 3.7 Feeding the Classifier

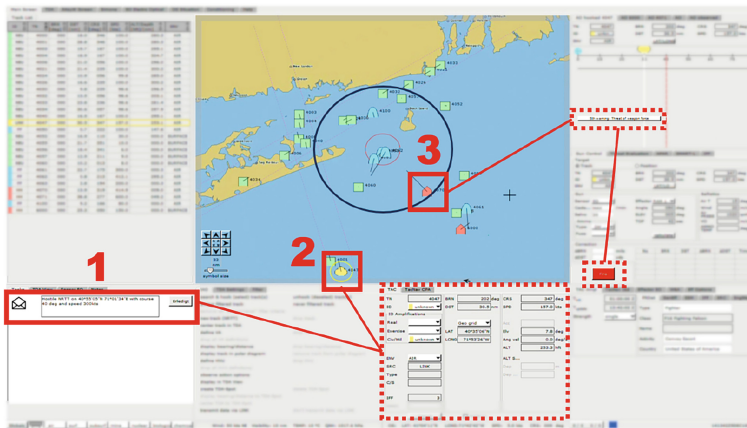
When a new slope has been calculated, we add the slope, its height and length to the previously normalized HRV values for the whole length of the calculated slope. For the length of the latest slope all normalized HRV values, the slope itself with its length and height will be passed to the classifier. For each normalized HRV value, the classifier will then predict if it is part of a mentally demanding sequence.

## 4 Artificial Intelligence Classifier Training

In this section, we describe how we trained the classifier, starting with the data we used, the model architecture, the training process, model evaluation, and finally further analysis.

### 4.1 Training Data and Preprocessing

For our study, we examined the video data of all participants from the study of Bruder and Schwarz [26] and annotated five different tasks. Participants were required to complete three surveillance tasks called NRTT, Unknown Track, and Warn/Engagement. Figure 9 shows the three different tasks and their associated areas.

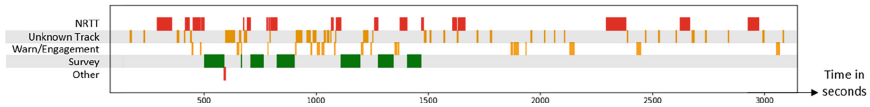


**Fig. 9.** Original Task User Interface, showing 1) a non-real-time track (NRTT) to process, 2) a new unknown track to classify and 3) a hostile track entering the self-protection zone that must be warned or further engaged if it proceeds to get closer.

The Non-Real-Time Tracks (NRTT) are displayed with information about a track that must be created manually, with specific information such as classification, speed, and direction, all of which must be entered into a form. Unknown tracks sometimes appeared as yellow symbols on the map, and the participant had to classify them within a form according to their position, speed, and direction. In the third task, hostile tracks,

marked as red symbols on the map, were moving towards our position in the center of the map. When these tracks crossed the first line, the participant had to manually warn the track by clicking a warn button to the right. If an enemy track crossed Fig. 10 on the right.

We annotated the surveys as a separate task and task-independent parts of the video, such as the start of the experimental software, as other. Visualizes all of a participant's tasks and their processing times. The three monitoring tasks could occur simultaneously which made the first half of the experiment mentally more demanding than the second half.



**Fig. 10.** Visualized time slots from one participant of all five different annotated classes.

We marked the three surveillance tasks as *mentally demanding tasks* which was coded as 1, whereas the periods surrounding these tasks were coded 0. The stated statistical model was defined as:

$$\text{Mentally demanding Task} \sim \text{normalized HRV} + \text{Slope} + \text{Length of Slope} + \text{Height of Slope} + \varepsilon$$

The calculation of the predictors as well as the criterion was based on the steps as described in paragraph 2.

We split the data into 70% training data and 30% test data for model validation afterwards. The training data consists of 70% of the dataset, which is around 6:52 h of training material, coded in 22.109 data rows. In contrast, the test data consists of 30% of the dataset, which is around 2:57 h of evaluation material, coded in 9.476 data rows that will only be used afterwards for the purpose of evaluation.

## 4.2 Model Architecture

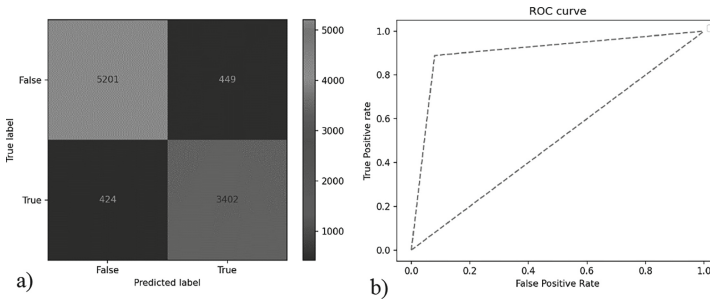
We have chosen a modified version of the Decision Tree Classifier algorithm called Extra Trees Classifier written in python v.3 from the package scikit-learn [27] in v.1.3.2. This package implements the Extra Trees Classifier as a meta estimator that fits several randomized decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and controls over-fitting, indicating its robust predictive capabilities in healthcare applications [28].

The hyperparameters were optimized using a grid search, resulting in a best fit with the parameters  $n\_estimators = 100$  and  $max\_features = 3$ .

With respect to possible class imbalance, we also calculated the class weights for the two possible states of the criterion, which were 0.839 for non-critical tasks and 1.236 for critical tasks. These class-weights were given to the following machine learning algorithm.

### 4.3 Training and Evaluation

Based on the test dataset, consisting of 30% (2:57 h) of the whole dataset, the classifier's discrimination accuracy is 90.78%, which is based on 8.603 right classified cases in contrast to 873 wrong classifications, as shown in the confusion matrix in Fig. 11.

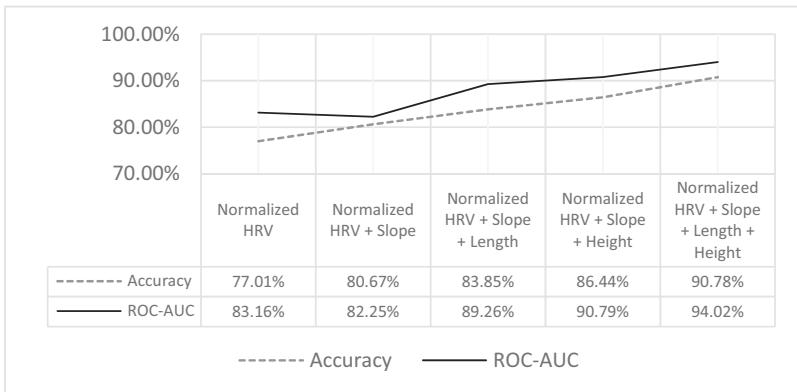


**Fig. 11.** a) Confusion Matrix based on test data. b) Receiver operating characteristic curve for the same test data.

A Receiver Operating Characteristic curve (ROC) is a graphical plot that illustrates the trade-off between the true positive rate (TPR) and the false positive rate (FPR) at different threshold settings, which is shown in Fig. 11 for our model evaluated with the test dataset. The calculated Receiver Operating Characteristic Area Under the Curve (ROC AUC) score indicates the ability of a model to distinguish between positive and negative examples across all possible classification thresholds. A ROC AUC score of 100% indicates a perfect model, while a score of 50% indicates a model that is no better than random guessing. Our trained model reaches an ROC AUC score of 94.02%, indicating that the model is very good in distinguishing between mentally demanding sequences in the HRV and sequences without. The ROC AUC score of our model is suggesting that the model is highly effective at identifying the true positive and true negative cases, also reflecting the results of the confusion matrix.

Feature importance measures the relative contribution of each feature to the classification accuracy of a machine learning model, which is a crucial aspect of understanding how a machine learning model works and can be used to improve its performance. It also shows how much information a specific feature adds to a model, which can help to decide whether the adding or deletion of a feature can optimize a model. In this study, we investigated the feature importance of the used predictors for predicting mental demanding sequences in the HRV. The four classes of features were the normalized HRV, the slope, its length and height. We found that the most important features are the normalized HRV the slope itself and the length of the slope, which are together accountable for 92.10% of the model's accuracy. This suggests that the height of an ascending or descending HRV, with a feature importance of 7.90%, is not as essential for predicting mental demanding tasks as the length of that slope with a feature importance of 26.96%. The slope itself, which is combining the length and height, has a feature importance valued with 26.49%. The normalized HRV has the highest feature importance with 38.65%.

Because of the partial redundancy of the predictors slope, length, and height, it may be sufficient to just use the slope and the normalized HRV values. In a further analysis, we trained a second model for testing with the normalized HRV and the slope as the only two predictors. Figure 12 shows the accuracies and ROC-AUC scores for different combinations of predictors.



**Fig. 12.** Accuracies and ROC-AUC scores for different combinations of predictors.

The further analysis indicates, that the normalized HRV explains most of the variance regarding mentally demanding tasks. Furthermore, the height of the slope seems to explain more variance than the length, but in the use of both features, variance of the height was lower as stated in the paragraphs before. In comparison to the model using just the normalized HRV + Slope the additional use of the predictors length or height increases the ROC-AUC score, therefore improving the model in identifying the true positive and true negative cases for up to ~8%.

In summary, our trained model detects mentally demanding tasks with a high accuracy of 90.78%. The ROC-AUC score of 94.02% shows that the model can distinguish positive and negative cases very well. The analysis shows that it appears possible to run the model with fewer predictors than features, thus reducing complexity.

## 5 Discussion

### 5.1 Benefits of Live HRV Analysis

Heart rate variability (HRV) analysis has long provided valuable insights into health and well-being. However, traditional methods often rely on establishing individual baselines, making them cumbersome and limiting.

Our live analysis reduces the complexity by tracking HRV fluctuations in near real time, enabling the identification of changes during mentally demanding tasks. This eliminates the need for lengthy baseline measurements, saving time in research settings and making the technology applicable outside of lab environments. Without spending time

establishing baselines, a system could start adapting to new situations faster, potentially leading to quicker and more optimal responses. Adapting without baselines could allow the system to learn and improve continuously, incorporating new information and experiences without needing to re-establish a static starting point.

In conclusion, our live HRV method indicates that real time classifications of short-time mentally demanding tasks based on HRV is possible. Additionally, our method eliminates the need for baseline measurements what offers significant advantages for real time applications like adaptive systems.

## 5.2 Limitations

Several limitations need to be considered regarding the overall use of HRV for measuring cognitive workload in general, and for the described method without a baseline specifically. One major limitation is that HRV is altered by many confounding variables, like body movement, general stress level, and psychoactive substances like caffeine. Discriminating whether a HRV change is caused, for example by physiological activity, homeostasis, or cognitive workload by analyzing the raw data, is difficult [29]. Context information are usually needed as a co-variable to address that issue. Those data are not always possible to collect. The applicability of HRV as a sole parameter to measure CWL is restricted because of that. To address the issue of discriminating different cofounding factors, the experiment was performed in a highly controlled laboratory environment. That limits the possible extrapolation of the results to field studies, or productive systems.

Another limitation is an increased data noise because the tasks for the training of the model were manually annotated, based on video material captured during the experiment.

Also, the described method was only used for a binary classification of cognitive workload, where multiple mentally demanding tasks at the same time were merged together. Usually, multiple tasks have to be monitored and performed independently, and some tasks are interrupted by others. The allocation of cognitive workload load due to specific tasks in that scenario has to be further investigated in future work and is a limitation of the results.

## 6 Conclusion and Future Developments

This study has demonstrated the potential of using heart rate variability (HRV) as a marker of cognitive workload, showing the way for further exploration and practical applications.

While this study successfully estimated mentally demanding tasks with an accuracy above 90%, future research can delve deeper into understanding the real time changes in HRV within single tasks or specific task types. This level of granularity could reveal which tasks or specific segments within tasks are most demanding, allowing for targeted interventions or workload balancing. Additionally, investigating the impact of multitasking on HRV could provide valuable insights into its unique cognitive demands.

Integrating other physiological metrics beyond HRV, such as pupil dilation, could potentially enhance the accuracy and comprehensiveness of workload assessment. Combining multiple physiological measures might create a more robust and nuanced understanding of cognitive state. While other physiological metrics, like pupil dilation, share

some similar connections to mental and emotional states as HRV, further research is needed to establish other metrics like the HRV as a reliable standalone measure for the intended method. It's important to note that replacing HRV with another physiological measure requires careful consideration of their specific functionalities and limitations within the given method's context.

Determining the optimal level of accuracy required for different applications is crucial. For instance, adaptive systems requiring real time adjustments might demand high accuracy, while workload evaluation might be tolerant of some margin of error. Tailoring the model's complexity and resource requirements to specific use cases will optimize its practicality and efficiency.

Future research could focus on differentiating mental and physical activities solely based on HRV data. This could enable applications like monitoring mental fatigue during physical exercise or distinguishing between cognitive stress and physical exertion in real time.

This study represents a step forward in utilizing HRV in live analysis to assess cognitive workload. Future research along the proposed avenues can refine and broaden this understanding, leading to impactful applications across various domains.

## References

1. Davis, D.H.J., Oliver, M., Byrne, A.J.: A novel method of measuring the mental workload of anaesthetists during simulated practice. *Br. J. Anaesth.* **103**(5), 665–669 (2009)
2. Galy, E., Paxion, J., Berthelon, C.: Measuring mental workload with the NASA-TLX needs to examine each dimension rather than relying on the global score: an example with driving. *Ergonomics* **61**(4), 517–527 (2018)
3. Mccraty, R., Shaffer, F.: Heart rate variability: new perspectives on physiological mechanisms, assessment of self-regulatory capacity, and health risk. *Global Adv. Health Med.* **4**(1), 46–61 (2015)
4. Kemper, K.J., Hamilton, C., Atkinson, M.: Heart rate variability: impact of differences in outlier identification and management strategies on common measures in three clinical populations. *Pediatr. Res.* **62**(3), 337–342 (2007)
5. Gąsior, J.S., et al.: Normative values for heart rate variability parameters in school-aged children: simple approach considering differences in average heart rate. *Front. Physiol.* **9**, 342109 (2018)
6. Li, W., Li, R., Xie, X., Chang, Y.: Evaluating mental workload during multitasking in simulated flight. *Brain Behav.* **12**(4), e2489 (2022)
7. John, A.R., et al.: Unravelling the physiological correlates of mental workload variations in tracking and collision prediction tasks: implications for air traffic controllers. *IEEE Trans. Neural Syst. Rehabil. Eng.* **30**, 770–781 (2021)
8. Delliaux, S., Delaforge, A., Deharo, J.C., Chaumet, G.: Mental workload alters heart rate variability, lowering non-linear dynamics. *Front. Physiol.* **10**, 565 (2019)
9. Radüntz, T., Mühlhausen, T., Freyer, M., Fürstenu, N., Meffert, B.: Cardiovascular biomarkers' inherent timescales in mental workload assessment during simulated air traffic control tasks. *Appl. Psychophysiol. Biofeedback* **46**, 43–59 (2020)
10. Forte, G., Favieri, F., Casagrande, M.: Heart rate variability and cognitive function: a systematic review. *Front. Neurosci.* **13**, 436204 (2019)
11. Veltman, J.A., Jansen, C.: Differentiation of mental effort measures: consequences for adaptive automation (2003)



12. De Witte, N.A., Sütterlin, S., Braet, C., Mueller, S.C.: Getting to the heart of emotion regulation in youth: the role of interoceptive sensitivity, heart rate variability, and parental psychopathology. *PLoS One* **11**, e0164615 (2016)
13. Veltman, J.A., Gaillard, A.W.K.: Indices of mental workload in a complex task environment. *Neuropsychobiology* **28**, 72–75 (1993)
14. Cinaz, B., Arnrich, B., Marca, R.L., Tröster, G.: Monitoring of mental workload levels during an everyday life office-work scenario. *Pers. Ubiquit. Comput.* **17**(2), 229–239 (2013)
15. Shao, S., Wang, T., Li, Y., Song, C., Jiang, Y., Yao, C.: Comparison analysis of different time-scale heart rate variability signals for mental workload assessment in human-robot interaction. *Wireless Commun. Mob. Comput.* **2021**, 1–12 (2021)
16. Nardolillo, A.M., Baghdadi, A., Cavuoto, L.A.: Heart rate variability during a simulated assembly task; influence of age and gender (2017)
17. Stern, R.M., Ray, W.J., Quigley, K.S.: Quigley, *Psychophysiological Recording*, vol. 59 (2001)
18. Ernst, G.: Methodological issues. In: *Heart Rate Variability*, pp. 51–118. Springer, London (2014)
19. Tarkiainen, T.H.: Stability over time of short-term heart rate variability. *Clin. Auton. Res.* **15**, 394–399 (2005)
20. Shaffer, F., Venner, J.: Heart rate variability anatomy and physiology. *Biofeedback (Online)* **41**, 13 (2013)
21. Chen, Z., Chen, Z., Calhoun, V.: Blood oxygenation level-dependent functional MRI signal turbulence caused by ultrahigh spatial resolution: numerical simulation and theoretical explanation. *NMR Biomed.* **26**(3), 248–264 (2013)
22. Oladele, A.M., Tomomowo-Ayodele, S.O., Oluremi, O.Y., Olusola, A.M.: Health information needs and its sources among rural dwellers in Egbedore local government areas of state of Osun, Nigeria. *Int. J. Humanit. Soc. Stud.* **7**(7) (2019)
23. Wilder, J.: Basimetric approach (law of initial value) to biological rhythms. *Ann. New York Acad. Sci.* **98**(4), 1211–1220 (1962)
24. Schwarz, J., Fuchs, S.: Validating a “Real-Time Assessment of Multidimensional User State” (RASMUS) for adaptive human-computer interaction (2018)
25. Hoover, A., Singh, A., Fishel-Brown, S., Muth, E.: Real-time detection of workload changes using heart rate variability. *Biomed. Signal Proc. Control* **7**, 333–341 (2012)
26. Bruder, A., Schwarz, J.: Evaluation of diagnostic rules for real-time assessment of mental workload within a dynamic adaptation framework. In: Sottilare, R., Schwarz, J. (eds.) *Adaptive Instructional Systems. Lecture Notes in Computer Science*(), vol. 11597. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-22341-0\\_31](https://doi.org/10.1007/978-3-030-22341-0_31)
27. Pedregosa, F., et al.: Scikit-learn: machine learning in Python (2011). ArXiv:abs/1201.0490
28. Gupta, M.D., et al.: COVID 19-related burnout among healthcare workers in India and ECG based predictive machine learning model: insights from the BRUCEE- Li study. *Indian Heart J.* **73**(6), 674–681 (2021)
29. Sammer, G.: Heart period variability and respiratory changes associated with physical and mental load: non-linear analysis. *Ergonomics* **41**(5), 746–755 (1998)
30. Bashiri, B., Mann, D.: Heart rate variability in response to task automation in agricultural semi-autonomous vehicles (2014)
31. Chamchad, D., et al.: Using heart rate variability to stratify risk of obstetric patients undergoing spinal Anesthesia (2004)