International Journal of Computing and Digital Systems
2025, VOL. 17, NO. 1, 1–17

http://dx.doi.org/10.12785/ijcds/1571111486

Optimizing Mental Workload Detection for HCI: Comparative Feature Selection and Interpretable Machine Learning

Auditya Purwandini Sutarto[1], Mega Bagus Herlambang[2], Nailul Izzah[1] and Ade Hendi[3]

[1]*Department of Industrial Engineering,Universitas Qomaruddin, Gresik, Indonesia*
[2]*Department of Industrial Engineering, Institut Teknologi Indonesia, Tangerang, Indonesia*
[3]*Department of Informatics Engineering,Universitas Qomaruddin, Gresik, Indonesia*

Received 21 September 2024, Revised 7 March 2025, Accepted 22 March 2025

Abstract: Detecting mental workload (MWL) is essential to optimize task performance, prevent cognitive overload, and improve well-being and safety, especially in contexts involving human-computer interaction (HCI) and complex task environments. This study introduces a novel data set for MWL detection using accessible physiological signals, specifically heart rate variability (HRV) and galvanic skin response (GSR), collected from 36 participants engaged in switch- and arithmetic tasks designed to induce varying levels of mental workload. To classify binary MWL states, we trained various machine learning (ML) algorithms, including Multi-Layer Perceptron (MLP), Gradient Boosting (GB), and Support Vector Machine (SVM). This approach provided significant insights for optimizing human-computer interaction (HCI) systems. To improve model performance, we employed three different feature selection techniques: correlation-based, minimum redundancy maximum relevancy (mRMR), and domain-expert selection. We optimized hyperparameters using grid search cross-validation (CV) and validated the results through nested CV. Among the models, the MLP with correlation-based feature selection demonstrated the highest performance, reaching an area under the curve (AUC) of 0.822. The GB and SVM models also performed well with mRMR and domain-expert feature selection (AUC = 0.760 and 0.741, respectively). To provide interpretability and a better understanding of feature importance in HCI contexts, Shapley Additive exPlanations (SHAP) identified several GSR features and heart rate as key predictors of mental workload, offering critical insights for designing adaptive HCI systems that account for cognitive load.

Keywords: Human-Computer Interaction, Feature Selection, Heart Rate Variability, Mental Workload, SHAP

1. Introduction

The workplace has undergone significant changes since the mid-20th century, with the rise of knowledge-based work and the increasing need for flexible work [1]. These changes highlight the importance of detecting mental workload (MWL) in preventing cognitive fatigue, decision-making errors, and productivity declines. Although no single definition of mental workload is universally agreed upon, a commonly accepted description by Young and Stanton defines it as the amount of attentional resources needed to meet performance requirements, affected by task complexity, available support, and previous experience [1]. MWL is often confused with cognitive load, but the two are distinct; cognitive load focuses on instructional design, while MWL emphasizes task demands and resource allocation [2]. MWL is commonly discussed in ergonomics and human-machine interaction and is classified into overload, underload, and optimal load states [3]. For this study, both terms will be used interchangeably; as in affective computing, the term ”cognitive load” is more widely used.

Advances in sensor technology and machine learning (ML) have enabled automatic detection of MWL in real-world environments, supporting self-monitoring and adaptive human-machine interaction (HMI) systems [4]. These systems adjust the workload to prevent overload or understimulation [5]. Physiological signals, such as heart rate variability (HRV) and Galvanic Skin Response (GSR), offer a reliable means of detecting cognitive load, as they reflect changes in the body’s autonomic nervous system [6], [7]. Wearable technologies now enable continuous, non-intrusive monitoring of these signals, making them invaluable for real-time MWL assessment.

While affective computing has been extensively studied, fewer publicly available datasets focus on mental workload compared to stress or emotion (for review, see [8], [9], [10], [11], [12]. Seitz et al. [13] reviewed cognitive load datasets, but few capture physiological signals from non-

E-mail address: audityapsutarto@gmail.com, mega.herlambang@iti.ac.id, nailul32@gmail.com, hendi.ade@gmail.com

intrusive devices. Therefore, new open datasets are essential for advancing research, reproducing results, and validating findings [14]. Since physiological signal correlations are not necessarily strong, ML models are crucial for reliably detecting workload, as they can account for individual differences in latent mental workload. Additionally, creating datasets tailored to specific populations, such as Indonesia, is important because physiological responses like HRV can vary by ethnicity [15].

On the other hand, enhancing the interpretation of features used in building ML models is critical for better understanding predictions and promoting fairness, trust, and transparency [16], [17], [18]. The importance of explainable ML (XML) is particularly emphasized in healthcare, where clinicians must be confident in AI systems to provide the best care for patients [19]. Healthcare poses unique ethical, legal, and regulatory challenges, as decisions can affect well-being or life immediately. However, these challenges also apply to fields like Human Factors and Ergonomics. AI systems in these domains, which are used to optimize workspaces, assess mental workload, or improve ergonomics, should also be transparent so practitioners can evaluate their impact on worker safety. This transparency builds trust among engineers and safety managers and ensures compliance with safety regulations. While ergonomic risks may not be as immediate as in healthcare, poor ergonomic designs can still have long-term effects on physical and mental health, highlighting the need for transparent, data-driven AI interventions to minimize potential harm.

One powerful XML technique used to enhance interpretability is SHapley Additive exPlanations (SHAP). SHAP, based on Shapley values from cooperative game theory, provides granular insights into how each feature contributes to a model's predictions, both individually and interactively [20]. It decomposes model outputs into additive contributions of features, making it possible to identify which variables drive predictions and how their values influence the model. By offering both global and local interpretability, SHAP helps ensure accountability and trust in ML systems, which is particularly important for safety-critical domains like ergonomics and mental workload assessment [21].

This transparency can be achieved by combining feature selection with explainable ML (XML) methods. When used together, feature selection and XML techniques provide a clearer understanding of how each feature contributes to predictions while reducing model complexity by removing irrelevant or redundant features. However, little research has been comparing domain expertise-driven and data-driven feature selection methods, particularly in HRV. Features should be chosen based on clinical or physiological motivations to enhance the contextual interpretation of model performance [16], [22]. Since HRV research offers well-established guidelines for key parameters [23], [24], comparing ML models using both selection methods is essential for detecting mental workload, and these should be evaluated alongside XML techniques.

Thus, this study aims to establish a new data set to detect mental workload using multimodal physiological signals and to evaluate the interpretability of ML models developed from this data set. It improves our prior study [25] by incorporating GSR data alongside HRV to enhance model performance. The data set focuses on the elicitation of mental workloads tailored to an Indonesian context. It also compares feature selection approaches—domain-expert and data-driven—in detecting MWL. To further enhance model interpretability, we performed SHapley Additive exPlanations (SHAP) to clarify the contribution of features and interactions differentiating MWL from non-load states [20]. The key contributions of this study include the following.

1) Introducing a novel dataset focused on the under-explored Indonesian population, improving classifier performance for adaptive HMI.
2) Comparison of domain-expert and data-driven feature selection in XML models, improving accountability, fairness, and transparency.
3) Developing an automatic mental workload detection system using accessible physiological signals from wearable devices, which offers potential for real-world mental workload monitoring.

2. Literature Review

A. Mental Workload Assessment from Physiological Data

Measuring MWL can be challenging, given its multi-faceted nature [1], [26]. Traditional methods have often relied on subjective self-reports or task performance metrics, but these approaches can be limited by their retrospective nature and the potential for bias. There has been a growing shift towards physiological measures, which offer objective and real-time insights into a user's cognitive load [26]. Recent advancements in sensor technologies have made physiological measurements more accessible, portable, and less intrusive. Tools now allow the collection of data such as HRV, electrodermal activity (EDA), respiration, electromyography (EMG), and photoplethysmography (PPG) without interrupting the primary task [26], [27]. These technologies enable researchers to gather precise and continuous data, which correlate directly with the user's mental state, offering a valuable alternative to traditional MWL assessments.

HRV and GSR, also known as EDA, are both valuable physiological indicators of mental workload, as they reflect the autonomic nervous system's (ANS) activity. The ANS regulates the body's involuntary physiological responses and consists of two branches: the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). The SNS is responsible for the "fight or flight" response during stress or high cognitive demand, while the PNS manages the "rest and digest" state, promoting relaxation.

HRV, the variation in time between heartbeats, is influenced by the balance between SNS and PNS activity [24]. Under increased mental workload, SNS activity rises, leading to an elevated heart rate and reduced HRV as the body becomes more alert and engaged with the task. Conversely, in more relaxed states, PNS activity dominates, resulting in lower heart rates and higher HRV, signaling a more adaptive and rested state. HRV can be derived from electrocardiography (ECG) and photoplethysmography (PPG), offering flexibility in physiological measurement.

HRV analysis can be performed in the time, frequency, and non-linear indices [24], [23], [28]. Time-domain measures assess the variability in the intervals between normal heartbeats (normal-to-normal intervals), frequency-domain measures quantify power distribution across different frequency bands using spectral analysis, and non-linear indices capture the complexity and unpredictability of heart rate time series.

Moreover, GSR, complements HRV as a physiological marker of ANS, specifically reflecting changes in SNS activation. GSR measures the skin's electrical conductivity, which increases with sweat gland activity driven by the SNS [29]. Since Skin Conductance Response (SCR) is primarily driven by SNS activity without significant influence from the PNS, it is an effective marker of arousal and stress. When cognitive demands or emotional arousal rise, the SNS activates sweat glands, particularly in the palms, increasing skin conductivity [7], [29].

In a mechanistic way, GSR captures variations in sweat-induced skin conductance, offering insights into emotional or cognitive states. It is a cost-effective and accessible physiological signal widely used in studies on mental workload, stress detection, and biofeedback interventions [7], [30].

GSR responses can be broadly classified into two distinct types: tonic and phasic [7], [29], [30]. Tonic responses refer to baseline levels of skin conductance measured during stimulus-free intervals, providing an indicator of general autonomic activity. Key features include the Skin Conductance Level (SCL), which represents the baseline conductance, and non-specific SCR (NS SCRs), which accounts for spontaneous fluctuations in skin conductance that occur without identifiable external stimuli.

In contrast, phasic responses reflect transient changes in skin conductance triggered by specific stimuli, offering valuable insights into immediate arousal or stress reactions. Phasic responses are further analyzed in two domains: the time domain and the frequency domain. Time domain features include measures such as Skin Conductance Response (SCR), peak amplitude, latency, and recovery time, while the frequency domain is assessed through metrics like SCR frequency. Table I summarizes the most commonly used HRV and GSR indices in psychophysiological research, along with their respective descriptions.

B. Related Work

Over the past decade, affective computing research has primarily focused on stress and emotional states, with relatively fewer studies dedicated to cognitive load inference. Understanding cognitive load is critical for developing adaptive HCI and human-machine interaction (HMI) systems that reduce distractions, minimize errors during demanding tasks, and enhance user performance. In the past five years, significant advancements have been made in detecting cognitive load using physiological signals and machine learning as summarized in Table II. Commonly used signals for workload detection include electrodermal activity (EDA), electrocardiography (ECG), and accelerometer (ACC) data.

As presented in Table II, multimodal approaches generally achieve higher accuracy rates. In classifying participants' expertise during a pedagogy simulation scenario, the accuracy of single-modality signals was 53.3% for ECG and 79.35% for GSR [34]. However, combining these signals using the kNN algorithm increased accuracy to 83.0%. Chalabianloo et al. [17] reported a similar trend. Although their work was excluded from the table because of its focus on stress detection devices, the authors found that incorporating a second biosignal (EDA) significantly enhanced performance, further supporting multisensor wearables' effectiveness. Moreover, most studies employed specific feature selection techniques, such as sequential backward floating selection (SBFS) [35], [36], [31] and minimum redundancy maximum relevancy (mRMR) [25], [32], [37], (see Table II. These techniques were typically applied alongside various k-fold cross-validation methods. However, none of the studies explicitly reported using hyperparameter optimization strategies, which could further enhance the robustness of their results.

Furthermore, there are public datasets that classify mental or cognitive load. Since we focused on long-term MW monitoring scenarios, we limited our summary to easily accessible physiological signals. Consequently, we do not consider brain-activity-related databases but instead concentrate on MWL public datasets that include easily acquired signals such as ECG, PPG, and GSR.

The datasets listed in Table III employ various strategies to induce various MW conditions, with the choice of stimuli being crucial. Different stimuli can evoke significantly different physiological responses or in some cases, no response. Thus, the selection of appropriate stimuli directly impacts their accuracy. For example, CLAS [38] used a series of interactive tasks like solving math problems, logic puzzles, and the Stroop test to generate different levels of MW. Participants were required to complete these tasks quickly to assess momentary cognitive load. In contrast, MMOD-COG [39] induced MW through various complex arithmetic tasks, with 4-operand arithmetic representing high complexity and 2-operand arithmetic representing low complexity. Similarly, CogLoad [35] and MAUS dataset

TABLE I. HRV and GSR Parameters and Their Descriptions

Parameter	Unit	Description
Heart Rate Variability (HRV)		
Time-domain		
Mean RR	ms	Mean of all RR intervals. The RR interval and heart rate (HR) are hyperbolically related (HR × RR interval = 60,000).
RMSSD	ms	Root mean square of successive differences between adjacent NN intervals (NN-intervals refer to interbeat intervals from which artifacts have been removed).
SDNN	ms	Standard deviation of NN intervals.
pNN50	%s	Percentage of successive NN intervals that differ by more than 50 ms.
Frequency-domain		
ULF power	ms^2	Absolute power of the ultra-low-frequency band (¡ 0.0033 Hz).
VLF power	ms^2	Absolute power of the very-low-frequency band (0.0033–0.04 Hz).
LF power	nu	Relative power of the low-frequency band (0.04–0.15 Hz) in normalized units.
HF power	nu	Relative power of the high-frequency band (0.15–0.4 Hz) in normalized units.
LF/HF	%	Ratio of LF to HF power.
Non-Linear		
SD1	ms	Poincaré plot standard deviation perpendicular the line of identity (short-term variability).
SD2	ms	Poincaré plot standard deviation along the line of identity (long-term variability).
SD2/SD1	%	Ratio of SD1 to SD2.
Galvanic Skin Response (GSR)		
SCL	**Microsiemens (μS)**	Baseline level of skin conductance during a stimulus-free interval.
SCR	μS	Phasic changes in skin conductance associated with a discrete stimulus.
Peak Amplitude	μS	The difference between the baseline and the highest point of the phasic response, indicating the intensity of the response.
Latency	Seconds	Time between the onset of a stimulus and the beginning of the SCR, measuring response speed.
Power Spectrum	Unitless (Normalized)	Frequency-domain feature calculated using Fourier Transform, often focusing on ranges below 1 Hz.
Impedance (SZ)	Ohms (Ω)	Resistance of the skin to alternating electrical current, reflecting electrodermal properties.
SCR Frequency	Count per minute	Number of discrete SCRs within a specified period, indicating arousal or stress level.

[40] used the 2-back and 3-back conditions, where higher N-back levels induced greater MW.

Signals recorded in these datasets include ECG, breathing rate, fingertip and wrist PPG, 3-axis ACC, GSR, and skin temperature. The datasets use different instruments for recording signals, resulting in varied sampling rates. In contrast, the CLAS and CogLoad datasets collected signals using wearable devices, and the MMOD-COG and MAUS datasets employed clinical-grade physiological monitoring systems. Furthermore, differences in MW stimuli and experimental setups lead to varying signal durations across the datasets. For example, the CLAS dataset provides 30-minute recordings per participant, while MMOD-COG offers 15-minute recordings. SVM was the most commonly used machine learning model in these public datasets, with accuracy rates ranging from 68.2% to 78.2%. Two datasets (MMOD-COG and CogLoad) used context (activity type) as ground truth, while CLAS relied on participant self-assessments, and MAUS employed NASA-TLX, a widely used self-report measure of mental workload [41].

3. Methods

A. Participants

A total of 36 healthy participants (16 males, 20 females, age = 21.9 ± 1.38 years, BMI = 23.22 ± 4.79) were recruited through advertisements on our university campus. Participants were required to avoid consuming caffeinated beverages for at least two hours before the experiment [42] and had a regular sleep routine the night before [43]. All participants were right-handed. To be eligible, they needed to meet the following criteria: 1) no history of

TABLE II. Overview of Related Work

Study & Year	N	Scenarios	Signals	Class	Feature Engineering	Ground Truth	Best Acc
[25], 2023	34	d2 attention, Switcher	ECG	Binary: MWL	FS: mRMR, domain expert, SFBS, NS Hype, LOOCV	Activity	67.89% SVM
[4], 2020	18	Arithmetic test	EMG, ECG, EDA	Multiclass: easy	FS: ANOVA, NS Hype, 10-fold	Activity	78.30% BPNN
[31], 2020	23	Maastricht Acute Stress Test	EOG, ECG	Binary: task	FS: SFFS, NS Hype, 8-fold CV	Activity	94.0% XGB
[32], 2019	24	Social exposure, event recall, cognitive load, stressful videos	ECG	Binary: mental stress	mRMR, NS Hyp., 10-fold CV	Activity	84.4% SVM
[33], 2019	16	Psychomotor vigilance task (PVT), *N*-back, visual search	ECG, EDA	Multiclass: types of task	NS FS, NS Hyp., LOOCV	Activity	66% kNN
[34], 2019	10	Trauma Simulation	ECG, GSR	Binary: novice and experts	LASSO, NS Hype, 5-fold CV	Expert	83.0% kNN

Notes. FS = Feature Selection, Hype = Hyperparameter Tuning, Acc = Accuracy, ECG= Electrocardiograph, EDA= Electrodermal Activity, EOG= Electrooculography, GSR= Galvanic Skin Response, NS= Not specified, FS= feature selection method, Hyp= Hyperparameter Tuning, LASSO= Least Absolute Shrinkage and Selection Operator, SFBS= Sequential Backward Floating Selection, SFFS= Sequential Forward Floating Selection, mRMR= minimum redundancy maximum Relevancy, SVM= Support Vector Machine, XGB= Extreme Gradient Boosting, BPNN: Back Propagation Neural Network.

TABLE III. Publicly Available Datasets from Related Work

Dataset (year)	N Subject	Scenario	Signals	Devices	Class (target)	Ground Truth	Best Acc
CLAS, 2019, [38]	59	Math & Logic Problem, Stroop Test, Neutral Music Video Clip	ECG, GSR, PPG	Shimmer 3	High and Low	Self-assessment	SVM 78.2%
MMOD-COG, 2019, [39]	40	Arithmetic Task	ECG, GSR, Speech	BIOPAC	Low and High	Context	SVM 76.66%
CogLoad, 2020, [35]	23	N-back Task	GSR, TEMP, ACC, ECG (PPG)	Microsoft Band	Rest and Task	Context	DT 68.2%
MAUS, 2021, [40]	22	N-back Task	ECG, PPG (Fingertip & Wrist), GSR, TEMP	Procomp Infiniti	High and Low Level	NASA-TLX	SVM 71.6%
Proposed dataset	36	Switcher Test, Arithmetic Task	ECG, EDA (GSR)	Polar H10, E-sense	MWL and not	Context	MLP 75.0%

Notes. ECG= Electrocardiograph, EDA= Electrodermal Activity, GSR= Galvanic Skin Response, PPG= Photoplethysmograph, TEMP= Skin Temperature, NS= Not specified, FS= Feature Selection Method, Hyp= Hyperparameter Tuning, mRMR= Minimum Redundancy Maximum Relevancy, SVM= Support Vector Machine, MLP= Multi-Layer Perceptron, DT= Decision Tree, NASA-TLX= NASA Task Load Index.

neurological, cardiac, or psychiatric disorders; 2) no long-term medical treatments; 3) smoking not more than four cigarettes per day; 4) body mass index (BMI) below 28 (obesity criteria); and 5) no allergies to adhesives or rubbing alcohol. Participation was voluntary, and written informed consent was obtained from each participant before the experiment. The protocol of the study's methodology adheres to the 1964 Declaration of Helsinki and obtained approval from the Health Research Ethics Committee (1497/KEP-UNISA/VII/2022).

B. Instruments and Devices

The sensors utilized in this study included the Polar H10 Heart Rate Sensor and the Mindfield eSense Skin Response Sensor for collecting physiological data. The Polar H10, a validated ECG chest strap, measures heart rate at 130 Hz and transmits RR intervals to a smartphone for real-time recording. GSR data were collected using the eSense Skin Response Sensor, which detects sweat gland activity through electrodes placed on the index and middle fingers. The skin conductance level and responses, measured in microSiemens (μS), was recorded as a time series on a smartphone at 5 Hz. While this study used the sensor's built-in output, future research could explore custom models for raw electrodermal data analysis.

C. Experimental Procedure

The experimental procedure began with an introduction where participants were informed about the data collection process, including its purpose and voluntary nature, with the option to withdraw at any time. After agreeing to participate, signing consent forms, and completing a demographics questionnaire, participants underwent the same experimental session, visiting the laboratory once.

Next, sensors were attached: the Polar H-10 electrocardiogram (ECG) on the chest and the Mindfield eSense GSR sensor on two fingers (index and middle) of the non-dominant hand. Baseline physiological responses were recorded for five minutes as participants sat upright in a relaxed position, avoiding movement, in line with the protocol [23]. This baseline measurement provided a reference

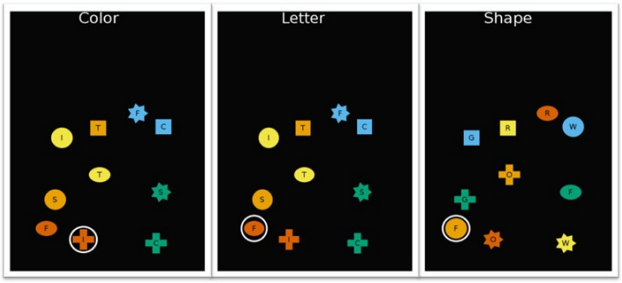

Figure 1. Switcher Featuring Tasks

point for assessing changes during subsequent tasks.

Participants then completed two cognitive tasks: a five-minute switcher task using PEBL software and a five-minute arithmetic task. The switcher task was implemented using the PEBL software. Participants were required to identify pairs of figure elements based on a switching feature rule among ten figures, each a unique combination of five colors, five shapes, and five letters displayed on a black screen (see Figure 1). Participants practiced briefly before the task, during which recording was paused. Following practice, participants completed nine blocks, each lasting until twelve correct answers were recorded. The task lasted approximately five minutes, and reaction time (RT) and accuracy were recorded for each trial [44], [45] (Figure 1).

The arithmetic task, part of the Trier Social Stress Test (TSST), required participants to count backward from 2043 in steps of seven under time pressure. All tasks followed well-documented and scripted protocols to ensure high data quality. Both the switcher and arithmetic tasks are widely accepted in the scientific community to induce autonomic nervous system related to MWL [44], [45], [46].

After completing both tasks, participants rested for five minutes (recovery). Next, the sensors were removed, and participants completed the NASA-TLX survey on perceived workload (not used in this analysis). The total session duration was about 30 minutes. Figure 2 illustrates the experimental procedure flow.

D. Machine Learning Pipeline

An illustration of the machine learning pipeline, including data pre-processing, feature selection, model selection, and evaluation, is shown in Figure 3; begins with signal and raw data pre-processing, moves through the ML pipeline, and concludes with explainable ML (XML) and performance metrics. All ML algorithms in this study were implemented using the Scikit-Learn Python package.

We implemented binary classification in this study. Given that the MWL—Switcher and Arithmetic tasks—were administered uniformly to all participants in a controlled laboratory setting, the known activity type (context) served as the data labels. This approach helps minimize potential bias that could arise from subjective self-reporting. Two authors were present throughout all data collection sessions, carefully noting each session's exact start and end times and the corresponding activity type. The experiment was structured to differentiate between various levels of mental workload and non-MWL conditions, including baseline and recovery ("no load") phases, as well as the Switcher and Arithmetic tasks ("load").

1) Data Preprocessing.
Data pre-processing followed guidelines from the North American Society for Pacing and Electrophysiology and the Task Force of the European Society of Cardiology [24], [23]. Since there are no formal guidelines for GSR recording duration, we matched it to the HRV window. Using session timestamps, raw heart rate data was divided into four phases: Baseline, MWL 1, MWL 2, and Recovery.
Following the recommendation, a 5-minute window was applied [24], [23]. The Elite HRV smartphone app extracted R-R intervals from the raw ECG data produced by the Polar H-10 [47]. Then, Kubios HRV (version 4.1.0), a scientifically validated software for HRV analysis, was used for artifact removal and HRV feature extraction [48]. It results in 24 HRV features across three domains: time (7 features), frequency (14 features), and non-linear (3 features). The final dataset contained 144 cases.
For GSR, we quantified skin conductance, which consists of two components: tonic (static level) and phasic (fluctuating responses), typically referred to as Skin Conductance Responses (SCR). A higher number of SCRs per minute generally indicates increased stress. Consistent with the scientific literature, a relaxed state is indicated by 0 to 5 SCRs per minute [49]. The device recorded several parameters, including the number of SCRs per minute, maximum and minimum values, the difference between min/max, total SCR, and the percentage of SCR per session.
We applied the MinMaxScaler from Scikit-learn during the pre-processing stage to perform min-max normalization. This technique adjusts each value by subtracting the feature's minimum and dividing the result by the feature's range (i.e., the difference between its maximum and minimum values). This process preserves the original distribution of the data without altering the underlying information in the features. By default, the MinMaxScaler scales values to a range between zero and one.
2) Feature Selection.
Feature selection is a process that aims to identify the optimal subset of features from a more extensive set that best differentiates between classes [50]. Reducing dimensionality through feature selection can improve classification accuracy. The performance of classifiers often depends on the features used. Feature selection methods are typically classified

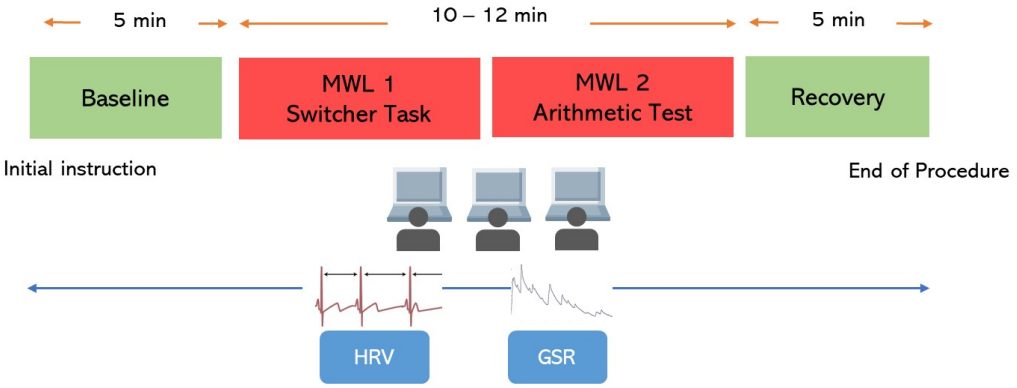

Figure 2. Experimental Procedure

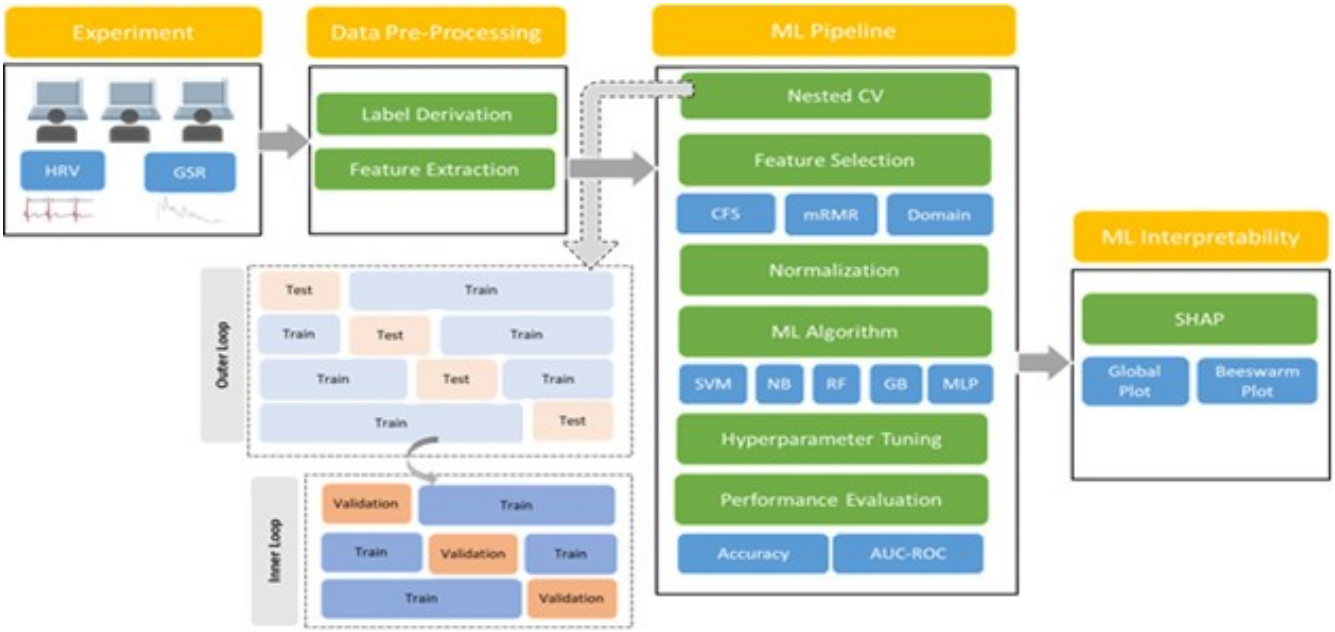

Figure 3. Steps of Our Methodology for the Mental Workload Detection with Physiological Signals and ML models

into three categories: filters, wrappers, and embedded methods [31], [51]. In this study, compare three methods of feature selections as follows:

- Domain Expertise.
  From the 24 HRV features extracted using Kubios HRV software (v4.1.0), we selected six based on recommendations from prior cognitive and HRV research [52], [23], [24]. Chosen features, both time and frequency domains, followed prior HRV and stress studies [53]. Key features include mean heart rate, SDNN, RMSSD, HF in normalized unit (HF nu), LF nu, and the ratio of LF to HF [24]. As to our knowledge, there is no recommendation from the profession or agreement on GSR; we relied on features provided by the internal system of the sensor manufacturer, so the total features used are six, including average SCR per minute, maximum value, minimum value, difference min/max, total SCR, and percentage of SCR per session.
- Correlation-based Feature Selection.
  This method evaluates feature subsets by considering their relevance to the target variable and their intercorrelation. Highly correlated features (Pearson correlation coefficient > 0.8) are considered redundant and are removed to retain the most relevant features. CFS is particularly effective when working with many features, as it helps identify the most important ones for the task.
- Minimum redundancy maximum relevancy (mRMR).
  This method selects features by considering

their relevance to the target variable and their redundancy with other selected features. The goal is to maximize target relevance while minimizing feature redundancy. Relevance is typically measured using F-statistics or mutual information, and redundancy is evaluated using Pearson correlation or mutual information. Features are ranked based on these criteria, and the most informative ones are selected iteratively [51].

Table IV shows the final of selected features

3) Cross Validation.
In the ML approach, cross-validation helps evaluate model performance on unseen data. Given the relatively small size of our dataset and the participant-dependent nature of the experimental sessions, we employed leave-one-out cross-validation (LOOCV) across all ML models. This CV type was also used in prior ML models for cognitive load and stress detection [9], [35], [54], [55]. Both hyperparameter optimization and model selection can be conducted through this procedure. However, using the exact data for hyperparameter tuning and model evaluation can introduce data leakage, leading to overfitting and overly optimistic results. To avoid this, we employed a nested cross-validation strategy, ensuring the reliability and fairness of the model performance assessment. The procedure was as follows:
   a) Outer Loop (Train-Test Split). Split the dataset into two subsets: 70% for training and 30% for testing. The test set is held out and will only be used for the final evaluation
   b) Inner Loop (Leave-One-Out Cross-Validation on Training Set).
      i) Within the training set (70% of the data), employ leave-one-out cross-validation (LOOCV).
      ii) In each iteration of LOOCV, one sample is held out as validation, while the remaining samples are used to train the model.
      iii) During this step, hyperparameter tuning is performed using methods like GridSearch to find the optimal parameters for the model
   c) Hyperparameter Optimization
      i) GridSearchCV is used within the inner loop to search for the best combination of hyperparameters across multiple iterations of LOOCV (see Table V). These parameters were selected based on their demonstrated effectiveness in prior research [56], [57], [58], empirical observations, and best practices for optimizing model performance in physiological-based workload detection.
      ii) After LOOCV is completed, the best model configuration (with optimized hyperparameters) is selected based on the performance in the inner loop.
   d) Model Evaluation with the Outer Loop
      i) Once the best hyperparameters are selected, the model is retrained on the entire training set (70%) using those parameters.
      ii) The model is then evaluated on the outer loop's held-out test set (30%) to assess its generalization performance on unseen data.

4) Classifiers (ML Algorithms).
The ML algorithms in this study were selected to compare classical techniques (SVM, kNN), ensemble methods (RF, GB), and Artificial Neural Networks (MLP) for detecting mental workload (MWL) based on our experimental data. Readers interested in further exploration of their applications and effectiveness in physiological signal analysis may refer to prior studies and comprehensive reviews (see [9], [10], [13].
   a) Support vector machines (SVM) are discriminative models designed to find an optimal hyperplane that separates data into different classes, particularly in high-dimensional spaces [59].
   b) Naïve Bayes is a family of probabilistic classifiers based on Bayes' theorem. It operates with strong independence assumptions between features and is computationally efficient due to its simple algorithmic structure. Naïve Bayes assumes that the value of each feature is independent of others within the same class [60].
   c) Random Forest (RF) is an ensemble learning method that combines multiple decision trees for accurate predictions [61]. It creates diverse trees by randomly sampling the dataset with replacements and makes final predictions by averaging or majority voting from these trees. RF is effective for classification and regression tasks, handling large datasets and reducing overfitting. Therefore, they are widely used in ML applications.
   d) Gradient Boosting (GB) is a popular ensemble machine learning technique that addresses the limitations of weak learners by iteratively improving their performance using gradient descent [62]. The loss function, which represents the difference between true and predicted values, is minimized with each step. By adding predictors sequentially, each new predictor corrects the errors of the previous one, thereby strengthening the overall model. GB is particularly effective in reducing the data's noise, variance, and bias, resulting in a more robust predictive model.

TABLE IV. List of the Features Selected

Method	HRV	GSR
Domain-expert	hr, sdnn, rmssd, lfnu, hfnu, lf/hf	'scr', 'max', 'diff_minmax', 'pscr'
mRMR	'minhr', 'maxhr', 'peak_lf', 'hfnu', 'hr', 'rr', 'lfnu'	scr, 'min', 'pscr'
CBS	'maxhr', 'peak_vlf', 'peak_lf', 'peak_hf', 'vlflog', 'hf', 'hfnu', 'lfhf', 'sd1', 'sd2', 'sd2sd1'	'scr', 'max', 'diff_minmax', 'pscr'

TABLE V. Predefined Hyperparameters for the GridSearch CV

Algorithm	Hyperparameter	Value
Support Vector Machine	C (regularization parameter)	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 1]
	gamma (kernel coefficient)	[1, 0.1, 0.01, 0.001, 0.0001]
	degree (for poly kernels)	[1, 2, 3, 4, 5]
	kernel	['linear', 'poly', 'rbf', 'sigmoid']
Naïve Bayes	Var smoothing	np.logspace(0, -9, num=100)
Random Forest	n_estimators	[200, 250, 300]
	criterion	['gini', 'entropy', 'log_loss']
	max_depth	[1, 2, 3, 4]
	min_samples_split	[2, 3]
	min_samples_leaf	[1, 2, 3]
GradientBoost	loss	['log_loss', 'exponential']
	learning_rate	[0.1, 0.3, 0.4, 0.5, 1.0, 1.1]
	n_estimators	[10, 20, 30, 50, 70]
Multi-Layer Perceptron	hidden_layer_sizes	[(100,), (200,), (300,), (400,)]
	activation	['identity', 'logistic', 'tanh', 'relu']
	solver	['sgd', 'adam']

e) Multi-layer perceptron (MLP) is a type of neural network that processes information in a forward direction, from input to output [63]. It consists of three types of layers: the Input Layer, Hidden Layer(s), and Output Layer. In MLP, each node, except in the Input Layer, represents a neuron that applies a non-linear activation function to transform the weighted sum of its inputs into an output. The Input Layer receives the input signal, while the desired regression or classification task is performed in the Output Layer.

5) Performance Metrics
In this study, machine learning performance was assessed using three primary metrics: accuracy, F1-score,and area under the curve (AUC). Accuracy, a basic metric, represents the ratio of correct predictions (true positives and true negatives) to the total number of predictions and is suitable for datasets with balanced class distributions, as in our case. However, for imbalanced datasets, accuracy can be misleading. To address this, we also utilized AUC-ROC (Area Under the Receiver Operating Characteristic Curve) and F1-score as complementary measures. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) for a binary classifier, and AUC quantifies the area under this curve. A perfect classifier has an AUC of 1.0, while an AUC of 0.5 indicates random guessing. Furthermore, F1 score provides a harmonic mean of precision and recall, ensuring that both false positives and false negatives are considered, ensuring practical effectiveness [63].

6) Estimation of SHAP Value
In this study, SHapley Additive exPlanations (SHAP) are utilized to enhance the interpretability of our machine learning models. SHAP is an open-source game-theoretic method that explains the prediction of the model by calculating the contribution of each feature to the final output. It relies on Shapley values from game theory, representing each feature's average marginal contribution in a coalition. SHAP generates explanations by simulating all possible combinations of features and measuring their individual contributions to the model's predictions. This approach quantifies each feature's positive or negative impact on the model's outcome, both on a global scale (across the entire dataset) and locally (for specific observations), thereby improving transparency in model interpretation [20].
The critical advantage of SHAP is its ability to explain any machine learning model with a set of feature contributions, offering more interpretable insights than standard feature importance scores. In

this study, SHAP was applied to our best-performing models, and two types of SHAP plots were generated—a global bar plot and a bee swarm plot—to provide deeper insights into the models' decision-making processes.

4. Results

A. Evaluation of ML Models

Table VI summarizes the performance of five machine learning algorithms (SVM, NB, GB, RF, and MLP) evaluated using three feature selection methods: CFS, domain expertise, and mRMR. The results reveal different patterns in how these models interact with the chosen feature selection methods.

With correlation-based feature selection, the MLP model consistently outperformed all others across metrics such as accuracy (0.7500), F1 score (0.7317), and AUC (0.8219). For mRMR feature selection, the SVM model performed best for accuracy 0.6818) and F1 score (0.6316) and GB achieved the highest AUC (0.7598). IIn contrast, the domain expertise method showed more variability across models. SVM performed best in this category, achieving the highest AUC (0.7412), suggesting that manually selected, relevant features improved the model's ability to classify effectively. RF achieved the highest accuracy (0.6818) and F1 scores (0.667). In general, MLP and SVM emerged as the models that performed the best, although their success depended on the feature selection method.

To complement the previously reported comparative metrics, we present confusion matrices for the MLP model with correlation-based feature selection and the SVM model with mRMR feature selection (Figure 4). These matrices, generated after hyperparameter tuning using a 70:30 training-to-test split, provide a more intuitive and visual representation of model performance.

B. Model Interpretation with SHAP Values

Figure 5 visualizes the contributions of the features of each feature selection method using the SHAP bar and the beeswarm plots for the tuned ML algorithms that perform the best. For CFS, the plots are based on MLP; for domain expertise, on SVM; and for mRMR, on GB. The SHAP bar plot displays the mean absolute SHAP values, highlighting the relative importance of each feature in the model's predictions, and longer bars indicate more influential features.

For Correlation-based feature selection, maximum GSR amplitude has the highest SHAP value (0.18), indicating its most influential feature, followed by 'sd1' (non-linear HRV feature, see Table I) and 'maxhr' (maximum heart rate). Features like 'sd2sd1' and 'LFHF' have less influence.

The SHAP beeswarm plot offers a comprehensive visualization of how each feature impacts the model's predictions across all data points. Features are ranked by importance along the y-axis, while the x-axis represents SHAP values, indicating the extent to which each feature increases or decreases the prediction. The color of each dot corresponds to the feature's value, with red signifying higher values and blue representing lower values. When multiple dots have the same SHAP value (x-axis position), they are stacked to represent the the density. The color scale on the right-hand side reflects the feature values, allowing for a quick visual assessment of how the feature magnitude affects the predictions. The plot highlights how feature values (represented by color) influence predictions, providing both global and local interpretability. For example, higher maximum GSR amplitude values are associated with positive predictions, while lower values have minimal features like 'diff_minmax', which show more scattered effects, indicating a less consistent impact. The plot highlights that 'max', 'sd1', 'maxhr', and 'sd2' are the most influential characteristics, driving predictions upward when their values are high. The beeswarm plot clarifies each feature's role and provides global and local insights into the model's decision-making.

The SHAP bar plot derived from the mRMR illustrates that 'percentage of skin conductance response ('pscr)) is the most influential feature with a SHAP value of 0.19, followed by an average of SCR ('scr') and heart rate ('hr) with values of 0.15 and 0.12, respectively. Lower-ranked features like maximum and minimum GSR amplitudes contribute less, with values around 0.03.

The SHAP beeswarm plot shows that 'pscr' and 'scr' appear as the most influential features, with higher values significantly increasing predictions, as shown by red dots to the right of zero. Lower-ranked features, such as 'minhr', 'rr', and 'hr', exhibit mixed effects with less consistent contributions. In summary, the SHAP bar and beeswarm plots emphasize the critical roles of 'pscr' and 'scr' in driving predictions, with other features like heart rate and HRV contributing variably to the model's decision-making process.

The SHAP bar plot, based on the domain expertise feature selection method, ranks 'pscr' (percentage of skin conductance response) as the most influential feature with a SHAP value of 0.19, followed by 'scr' and 'hr', with values of 0.15 and 0.12, respectively. Lower-ranked features, such as 'max' and 'min' (GSR amplitudes), contribute less with SHAP values around 0.03. Skin conductance and heart rate measures, particularly 'pscr', play the most critical roles in driving predictions.

The SHAP beeswarm plot complements this by showing how individual data points contribute to the model's predictions. Features such as 'pscr', 'scr', and 'hr' again emerge as the most important, with higher values (in red) pushing predictions upward. HRV metrics (like 'lfnu' and 'hfnu') and 'diff_minmax' have more variable and inconsistent impacts, while features like 'max' and 'min' show weaker effects. In summary, 'pscr', 'scr', and 'hr' are

TABLE VI. Machine Learning Evaluation Results with Binary Classification Accuracy

	CFS			mRMR			Domain Expertise		
	Acc	**AUC**	**F1**	**Acc**	**AUC**	**F1**	**Acc**	**AUC**	**F1**
SVM	0.6591	0.707	0.6512	**0.6818**	0.6149	**0.6316**	0.6364	**0.7412**	0.6190
Naïve	0.5909	0.6128	0.5714	0.5682	0.6377	0.5366	0.6591	0.7019	0.6512
GB	0.6818	0.6812	0.500	0.6364	**0.7598**	0.5854	0.6818	0.6708	0.6500
RF	0.5909	0.6563	0.6316	0.6136	0.6749	0.6000	**0.6818**	0.6584	**0.6667**
MLP	**0.7500**	**0.8219**	**0.7317**	0.5682	0.6853	0.5581	0.5909	0.6439	0.5909

Notes. CFS= Correlation Feature Selection, mRMR= minimum redundancy maximum relevancy. Best performance based on accuracy, F1 scores, and AUC are marked in Bold.

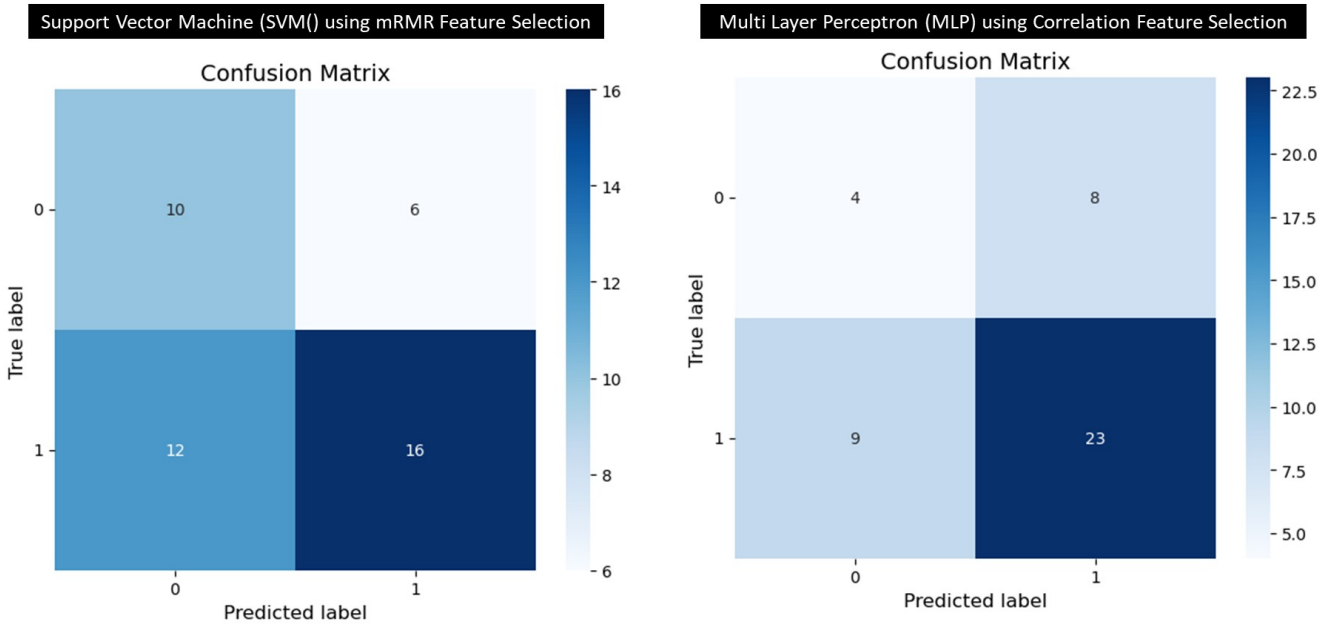

Figure 4. Confusion Matrices for SVM using mRMR and MLP using correlation feature selection

important, while other physiological metrics contribute less consistently to the model's predictions

5. Discussion

In this study, we present a new dataset and develop a machine-learning pipeline to detect mental workload using physiological data from ECG (HRV) and GSR, collected in a laboratory experiment with 36 participants across two types of MWL tasks. The pipeline integrates various algorithms, including SVM, kNN, RF, GB, and MLP. It also incorporates correlation, mRMR, and domain expert-based feature selection, nested CV, hyperparameter tuning, and SHAP value computation for model interpretability.

Our findings show that the MLP model, trained on HRV and GSR signals, outperformed the other models in predicting mental workload, achieving the highest accuracy and AUC scores. GB performed best with the mRMR method, while SVM yielded the best results with domain-expertise-based feature selection. In general, MLP and SVM emerged as the top-performing models, although their success depended on the feature selection method.

Overall, data-driven feature selection methods (correlation and mRMR) provided more consistent and stable results than the knowledge-domain approach. Correlation-based methods strongly favored MLP, while mRMR demonstrated synergy with SVM and RF models. The knowledge-domain approach, while applicable, showed more variability and lacked the consistency observed with the data-driven techniques. These results highlight the importance of selecting appropriate feature selection techniques tailored to specific ML models to optimize performance.

A. *Comparison with Other Public Dataset*

Directly comparing our study with previous works is challenging due to substantial differences in experimental setups, including variations in datasets, sensors, preprocessing techniques, machine learning methods, classification tasks, and evaluation procedures. Moreover, discrepancies in the generated features and performance metrics make such comparisons even more difficult. While some prior studies report high performance in detecting cognitive or mental load, these results may be overly optimistic due to methodological limitations. Despite these constraints,

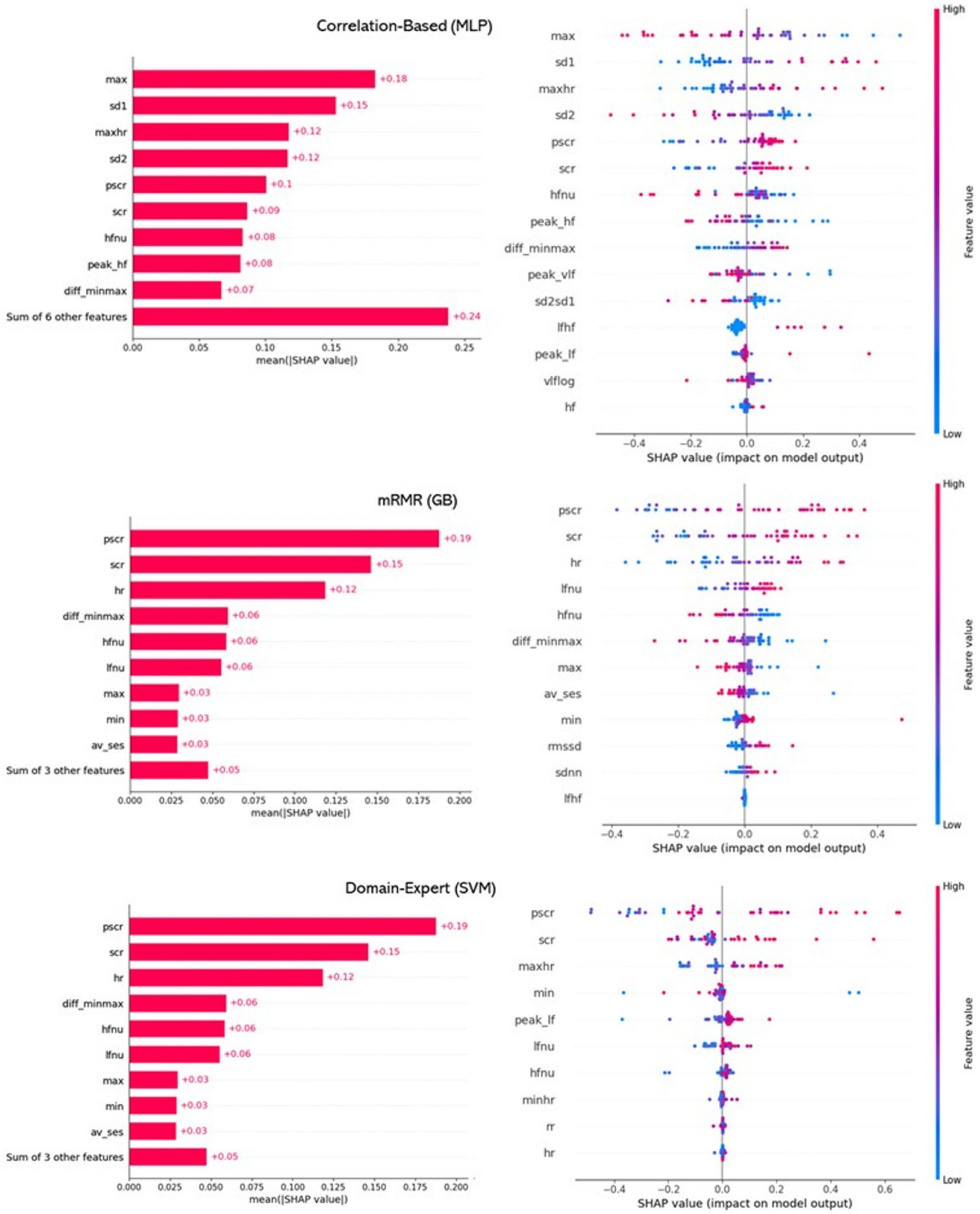

Figure 5. Global SHAP Plot (Left) and Beeswarm SHAP Plot (Right) for the Best Model from Each Feature Selection Method

we provide a comparison with selected notable studies, particularly those listed in Table III, to offer context for our findings. Regarding the experimental protocol used to generate training data, the work by [35] may be the most comparable to ours. Their experiment also aimed to induce cognitive load, with 23 participants completing 2-back and 3-back tasks. For ML model development, the authors utilized a feature selection ranking method based on mutual information and LOOCV. Their study employed various algorithms, RF, NB, KNN, Logistic Regression (LR), Decision Trees (DT), GB, XGB, and MLP, to distinguish the cognitive load from non-load from GSR and HRV features. Their best performance was achieved with DT (68.2%), while other algorithms, such as RF (67.9%), NB (60.8%), and MLP (64.3%), showed results similar to ours. Moreover, as shown in Table III, CLAS showed the highest accuracy of ML models for the public dataset, which resulted in 78.2% using SVM to detect high and low cognitive load using ECG and GSR signals [38]. However, this data employed various stimuli (mathematics and logic problems, Stroop test, neutral music, and video clip), administered in less than five minutes. These might not be enough to induce a mental workload.

B. Comparison across Feature Selection and Algorithms

Our results indicate that MLP, a neural network-based ML algorithm, outperformed other methods in a binary classification of mental workload data using correlation-based feature selection (CFS), hyperparameter optimisation, and nested cross-validation. Despite the relatively small dataset, this strong performance can be attributed to CFS, which reduces dimensionality by selecting only the most relevant features, helping to prevent overfitting. By removing irrelevant or redundant features, MLP can focus on the most critical patterns in the data. MLP's capacity to model complex, non-linear relationships through its hidden layers and regularization techniques, such as dropout, further enhances its generalization ability. The superiority of MLP was also demonstrated in prior studies using a larger number of instances [35], [64], [17].

With regard to domain expertise feature selection, the superior performance of SVM in our study might be attributed to the linear nature of the selected features, especially those related to HRV. These features exhibit linear properties, which align well with SVM's strength in handling linearly separable data. Additionally, SVM is well-suited for small datasets like our datasets, as it maximizes the margin between classes. Our HRV data appears to have a distinct margin of separation, making SVM efficient and accurate in this context.

However, SVM tends to scale poorly with larger datasets, potentially becoming computationally expensive in future studies involving more subjects or additional physiological signals and cognitive states. This observation supports a work by Choi et al. [54], which noted that SVM often demonstrates superior generalization ability, mainly when training data is limited and across diverse samples [34], [35], [31].

C. Interpretation of Results with SHAP Values

The SHAP analysis across the three feature selection methods (CFS, mRMR, and domain expertise) consistently highlights the significance of specific physiological features, though their relative influence varies depending on the selection method. Across all methods, 'pscr' (percentage of SCR) and 'scr' (average skin conductance response) consistently emerge as the most influential features, emphasizing the critical role of skin conductance in detecting mental workload. Skin conductance, which reflects changes in sweat gland activity driven by autonomic nervous system arousal, serves as a sensitive indicator of physiological stress or workload, making it central to model predictions.

Heart rate ('hr') also appears as a key feature in all selection methods, further validating its role in workload detection. It provides valuable insights into mental and emotional states, complementing the strong influence of skin conductance ('pscr' and 'scr'). These findings aligned with prior studies that demonstrate the effectiveness of combining electrodermal activity and cardiac signals for mental state classification [38], [35], [39], [40].

In contrast, features related to HRV and GSR amplitudes ('max', 'min') exhibited consistent influence across models. While HRV metrics such as 'hfnu', 'lfnu', and 'LF/HF ratio' are known to capture ANS dynamics [23], [52], their predictive utility appears more context-dependent. This inconsistency may be attributed to the individual differences, environmental conditions, or measurement noise. Similarly, GSR amplitude features like 'max' and 'min' may provide supplementary information but lack the direct and robust influence observed with 'pscr' and 'scr'.

Feature selection also significantly affected the interpretability of the model. For example, CFS prioritized 'maxhr' and 'sd1', while mRMR and domain expertise favored skin conductance metrics. This highlights the importance of selecting features tailored to specific contexts.

Furthermore, the SHAP beeswarm plots provide valuable insights into feature contributions at the individual data point level. For example, higher values of 'pscr' and 'scr' (red dots) are strongly associated with positive predictions for mental workload, while lower values (blue dots) contribute less to predictions. This local interpretability complements global feature rankings and underlines the importance of both individual feature values and their interactions in driving predictions.

These findings highlight the need for further investigation into the contextual relevance of features. Although skin conductance metrics are consistently dominant, HRV features may play a more critical role in specific conditions. Future research should integrate domain expertise with data-driven approaches to refine feature selection and improve

model adaptability across diverse populations and contexts.

D. Practical Implications for Mental Workload Detection, Workplace Applications, and Human-Computer Interaction

The findings of this study have several implications for the detection of mental workload in various human-computer interaction (HCI) scenarios. By demonstrating high accuracy through HRV and GSR, in particular, SCR features, this study shows that cost-effective, wearable solutions can provide continuous and unobtrusive insights into user cognitive states.

In office environments, for example, adaptive interfaces could automatically reduce visual complexity or pause notifications when MWL rises, preventing burnout and improving efficiency. In surgical settings, unobtrusive monitoring of GSR and HRV can provide an early warning system for surgeon fatigue. This data can then be used to implement proactive strategies, such as targeted rest breaks or reallocation of surgical tasks, potentially improving patient outcomes. Furthermore, in supervisory process control, where operators manage critical systems (such as power plants or chemical facilities), a reliable system for detecting MWL could proactively prevent errors by alerting operators to impending cognitive overload.

The study also underlines the importance of tailored feature selection strategies to optimize performance across different machine learning algorithms. This flexibility allows HCI designers and developers to choose the most suitable methods for real-time monitoring in diverse environments, ranging from high-pressure scenarios like surgeries to everyday contexts. This could involve employing MLP with correlation-based feature selection or SVM with domain-expert-defined features. Furthermore, SHAP-based explainability provides insights into the reasons behind specific interface adaptations. This is crucial for building user trust and ensuring regulatory compliance, particularly in safety-critical fields such as surgery and industrial control.

Lastly, while laboratory-based validations demonstrate feasibility, field studies are crucial. Evaluating these models in real-world environments, such as busy offices, high-pressure operating rooms, and industrial control stations, will determine their robustness, inform practical interface adaptations, and determine how to integrate them into existing workflows. Future research could expand the sensor suite to include EEG, pupillometry, or EMG signals, potentially improving model accuracy and broadening their applicability across various HCI scenarios. This multi-faceted approach will ensure that MWL detection systems remain responsive, reliable, and user-centered.

E. Limitations and Future Research Direction

While this research shows the potential to predict mental workload (MW) using easily acquired physiological signals, several limitations should be noted. First, the small and homogeneous sample of young university students limits generalizability. Future studies should involve larger, more diverse samples. Second, although we used the recommended 5-minute window segments [23], [24], shorter windows (30–60 seconds), commonly used in HRV-based studies [64], could increase data points, allow the classification of MWL levels by multiple classes and the use of deep learning models. Third, incorporating simulated tasks that resemble real-life scenarios would improve external validity. Including subjective self-reports, as done in other studies [4], [16], [35], would also offer insights into participants' perceived stress or relaxation levels, providing a fuller understanding of MW as both an input and an output [26].

Multimodal data integration, such as eye movements, facial expressions, and task performance, is recommended for a more comprehensive analysis. Lastly, while we used two separate devices (Polar H10 and eSense) to collect signals, future studies could benefit from a single wearable device such as the wrist-worn Empatica E4, which, though less precise, would be more practical and less intrusive for field research.

6. Conclusion

The cognitive or mental workload of operators is crucial for optimizing human resource allocation, improving task performance, and ensuring well-being and safety. This study introduces a new dataset that predicts mental workload from easily acquired physiological signals, specifically HRV and GSR. Data were collected in a laboratory setting where participants completed switch and arithmetic tasks designed to induce mental workload. The study also explores various feature selection techniques, extensive hyperparameter tuning, and machine learning algorithms to classify resting versus workload states based on HRV and GSR features. Our results show that the MLP algorithm achieved the best performance for MWL detection with feature selection, while Gradient Boosting performed best with mRMR and SVM, which excelled using domain knowledge. Additionally, using SHAP values as an explainable ML method demonstrated the dominant role of skin conductance and heart rate in driving model predictions. Features like HRV metrics and GSR amplitudes exhibited secondary and context-dependent contributions. Future research should include field studies to validate these findings in real-world conditions, such as office work or monitoring tasks with actual employees. Such studies would enhance the external validity of the research and contribute to the development of adaptive, personalized systems for continuous mental workload detection in workplace environments.

7. Acknowledgment

This work was supported by the Directorate of Indonesian Higher Education under Research Grant for Junior Lectures [Grant number: 068/SP2H/PT/LL7/2024]

References

[1] M. S. Young, K. A. Brookhuis, C. D. Wickens, and P. A. Hancock, "State of science: mental workload in ergonomics," *Ergonomics*, vol. 58, no. 1, pp. 1–17, Jan. 2015. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00140139.2014.956151

[2] T. Luong, A. Lecuyer, N. Martin, and F. Argelaguet, "A Survey on Affective and Cognitive VR," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 12, pp. 5154–5171, Dec. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9531381/

[3] R. McKendrick, B. Feest, A. Harwood, and B. Falcone, "Theories and Methods for Labeling Cognitive Workload: Classification and Transfer Learning," *Frontiers in Human Neuroscience*, vol. 13, no. September, pp. 1–20, 2019.

[4] Y. Ding, Y. Cao, V. G. Duffy, Y. Wang, and X. Zhang, "Measurement and identification of mental workload during simulated computer tasks with multimodal methods and machine learning," *Ergonomics*, vol. 63, no. 7, pp. 896–908, Jul. 2020. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/00140139.2020.1759699

[5] V. Borisov, E. Kasneci, and G. Kasneci, "Robust cognitive load detection from wrist-band sensors," *Computers in Human Behavior Reports*, vol. 4, p. 100116, Aug. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2451958821000646

[6] T. Pereira, P. R. Almeida, J. P. Cunha, and A. Aguiar, "Heart rate variability metrics for fine-grained stress level assessment," *Computer Methods and Programs in Biomedicine*, vol. 148, pp. 71–80, 2017.

[7] F. Chen, J. Zhou, Y. Wang, K. Yu, S. Z. Arshad, A. Khawaji, and D. Conway, "Galvanic Skin Response-Based Measures," in *Robust Multimodal Cognitive Load Measurement*. Cham: Springer International Publishing, 2016, pp. 87–99, series Title: Human–Computer Interaction Series. [Online]. Available: http://link.springer.com/10.1007/978-3-319-31700-7_5

[8] P. J. Bota, C. Wang, A. L. Fred, and H. Placido Da Silva, "A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals," *IEEE Access*, vol. 7, pp. 140 990–141 020, 2019.

[9] S. Gedam and S. Paul, "A Review on Mental Stress Detection Using Wearable Sensors and Machine Learning Techniques," *IEEE Access*, vol. 9, pp. 84 045–84 066, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9445082/

[10] S. S. Panicker and P. Gayathri, "A survey of machine learning techniques in physiology based mental stress detection systems," *Biocybernetics and Biomedical Engineering*, vol. 39, no. 2, pp. 444–469, 2019, publisher: Nalecz Institute of Biocybernetics and Biomedical Engineering of the Polish Academy of Sciences. [Online]. Available: https://doi.org/10.1016/j.bbe.2019.01.004

[11] P. Schmidt, A. Reiss, R. Dürichen, and K. V. Laerhoven, "Wearable-based affect recognition—a review," *Sensors (Switzerland)*, vol. 19, no. 19, pp. 1–42, 2019.

[12] G. Vos, K. Trinh, Z. Sarnyai, and M. Rahimi Azghadi, "Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review," *International Journal of Medical Informatics*, vol. 173, no. February, p. 105026, 2023, publisher: Elsevier B.V. [Online]. Available: https://doi.org/10.1016/j.ijmedinf.2023.105026

[13] J. Seitz and A. Maedche, "Biosignal-Based Recognition of Cognitive Load: A Systematic Review of Public Datasets and Classifiers," in *Information Systems and Neuroscience*, F. D. Davis, R. Riedl, J. Vom Brocke, P.-M. Léger, A. B. Randolph, and G. R. Müller-Putz, Eds. Cham: Springer International Publishing, 2022, vol. 58, pp. 35–52, series Title: Lecture Notes in Information Systems and Organisation. [Online]. Available: https://link.springer.com/10.1007/978-3-031-13064-9_4

[14] L. Longo, C. D. Wickens, G. Hancock, and P. A. Hancock, "Human Mental Workload: A Survey and a Novel Inclusive Definition," *Frontiers in Psychology*, vol. 13, p. 883321, Jun. 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.883321/full

[15] L. K. Hill, D. D. Hu, J. Koenig, J. J. Sollers, G. Kapuku, X. Wang, H. Snieder, and J. F. Thayer, "Ethnic differences in resting heart rate variability: A systematic review and meta-analysis," *Psychosomatic Medicine*, vol. 77, no. 1, pp. 16–25, 2015, iSBN: 0000000000000.

[16] M. Bahameish, T. Stockman, and J. Requena Carrión, "Strategies for Reliable Stress Recognition: A Machine Learning Approach Using Heart Rate Variability Features," *Sensors*, vol. 24, no. 10, p. 3210, May 2024. [Online]. Available: https://www.mdpi.com/1424-8220/24/10/3210

[17] N. Chalabianloo, Y. S. Can, M. Umair, C. Sas, and C. Ersoy, "Application level performance evaluation of wearable devices for stress classification with explainable AI," *Pervasive and Mobile Computing*, vol. 87, p. 101703, Dec. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S157411922200116X

[18] J. S. Banerjee, M. Mahmud, and D. Brown, "Heart Rate Variability-Based Mental Stress Detection: An Explainable Machine Learning Approach," *SN Computer Science*, vol. 4, no. 2, p. 176, Jan. 2023. [Online]. Available: https://link.springer.com/10.1007/s42979-022-01605-z

[19] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies," *Journal of Biomedical Informatics*, vol. 113, p. 103655, Jan. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1532046420302835

[20] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, Jan. 2020. [Online]. Available: https://www.nature.com/articles/s42256-019-0138-9

[21] F. Cabitza and A. Campagner, "The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies," *International Journal of Medical Informatics*, vol. 153, no. xxxx, p. 104510, 2021, publisher: Elsevier B.V. [Online]. Available: https://doi.org/10.1016/j.ijmedinf.2021.104510

[22] B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," *Computers in Biology and Medicine*, vol. 112, p. 103375, Sep. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0010482519302525

[23] S. Laborde, E. Mosley, and J. F. Thayer, "Heart rate variability and cardiac vagal tone in psychophysiological research - Recommendations for experiment planning, data analysis, and data reporting," *Frontiers in Psychology*, vol. 8, no. 213, 2017.

[24] M. Malik, "Heart rate variability Standards of measurement, physiological interpretation, and clinical use," *European Heart Journal*, vol. 17, no. 3, pp. 354–381, 1996, iSBN: 9781482243482.

[25] N. Izzah, A. P. Sutarto, A. Hendi, M. Ainiyah, and M. N. B. Abdul Wahab, "Physiological Signals as Predictors of Mental Workload: Evaluating Single Classifier and Ensemble Learning Models," *Jurnal Optimasi Sistem Industri*, vol. 22, no. 2, pp. 81–98, Dec. 2023. [Online]. Available: https://josi.ft.unand.ac.id/index.php/josi/article/view/677

[26] R. L. Charles and J. Nixon, "Measuring mental workload using physiological measures: A systematic review," *Applied Ergonomics*, vol. 74, no. September 2016, pp. 221–232, 2019, publisher: Elsevier. [Online]. Available: https://doi.org/10.1016/j.apergo.2018.08.028

[27] D. Tao, H. Tan, H. Wang, X. Zhang, X. Qu, and T. Zhang, "A systematic review of physiological measures of mental workload," *International Journal of Environmental Research and Public Health*, vol. 16, no. 15, pp. 1–23, 2019.

[28] M. Mohanty, P. Sekhara Rath, and A. G. Mohapatra, "IoMT-based Heart Rate Variability Analysis with Passive FBGSensors for Improved Health Monitoring," *International Journal of Computing and Digital Systems*, vol. 15, no. 1, pp. 1135–1147, Mar. 2024. [Online]. Available: https://journals.uob.edu.bh/handle/123456789/5201

[29] W. Boucsein, *Electrodermal Activity*. Boston, MA: Springer US, 2012.

[30] N. Nourbakhsh, Y. Wang, F. Chen, and R. A. Calvo, "Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks," in *Proceedings of the 24th Australian Computer-Human Interaction Conference*. Melbourne Australia: ACM, Nov. 2012, pp. 420–423.

[31] K. Pettersson, J. Tervonen, J. Narvainen, P. Henttonen, I. Maattanen, and J. Mantyjarvi, "Selecting Feature Sets and Comparing Classification Methods for Cognitive State Estimation," *Proceedings - IEEE 20th International Conference on Bioinformatics and Bioengineering, BIBE 2020*, pp. 683–690, 2020, iSBN: 9781728195742.

[32] G. Giannakakis, K. Marias, and M. Tsiknakis, "A stress recognition system using HRV parameters and machine learning techniques," *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2019*, pp. 269–272, 2019, publisher: IEEE ISBN: 9781728138916.

[33] H. F. Posada-Quintero and J. B. Bolkhovsky, "Machine learning models for the identification of cognitive tasks using autonomic reactions from heart rate variability and electrodermal activity," *Behavioral Sciences*, vol. 9, no. 4, 2019.

[34] K. Ross, P. Sarkar, D. Rodenburg, A. Ruberto, P. Hungler, A. Szulewski, D. Howes, and A. Etemad, "Toward Dynamically Adaptive Simulation: Multimodal Classification of User Expertise Using Wearable Devices," *Sensors*, vol. 19, no. 19, p. 4270, Oct. 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/19/4270

[35] M. Gjoreski, T. Kolenik, T. Knez, M. Luštrek, M. Gams, H. Gjoreski, and V. Pejović, "Datasets for cognitive load inference using wearable sensors and psychological traits," *Applied Sciences (Switzerland)*, vol. 10, no. 11, 2020.

[36] C. Y. Chen, C. J. Wang, E. L. Chen, C. K. Wu, Y. K. Yang, J. S. Wang, and P. C. Chung, "Detecting sustained attention during cognitive work using heart rate variability," *Proceedings - 2010 6th International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIHMSP 2010*, pp. 372–375, 2010, iSBN: 9780769542225.

[37] K. Dahal, B. Bogue-Jimenez, and A. Doblas, "Global Stress Detection Framework Combining a Reduced Set of HRV Features and Random Forest Model," *Sensors*, vol. 23, no. 11, p. 5220, May 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/11/5220

[38] V. Markova, T. Ganchev, and K. Kalinkov, "CLAS: A Database for Cognitive Load, Affect and Stress Recognition," in *2019 International Conference on Biomedical Innovations and Applications (BIA)*. Varna, Bulgaria: IEEE, Nov. 2019, pp. 1–4.

[39] I. Mijic, M. Sarlija, and D. Petrinovic, "MMOD-COG: A Database for Multimodal Cognitive Load Classification," in *2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA)*. Dubrovnik, Croatia: IEEE, Sep. 2019, pp. 15–20.

[40] W.-K. Beh and Y.-H. Wu, "MAUS: A Dataset for Mental Workload Assessment on N-back Task Using Wearable Sensor," *Arxiv*, 2021.

[41] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," *Advances in Psychology*, vol. 52, no. C, pp. 139–183, 1988.

[42] F. Zimmermann-Viehoff, J. Thayer, J. Koenig, C. Herrmann, C. S. Weber, and H.-C. Deter, "Short-term effects of espresso coffee on heart rate variability and blood pressure in habitual and non-habitual coffee consumers – A randomized crossover study," *Nutritional Neuroscience*, vol. 19, no. 4, pp. 169–175, Apr. 2016. [Online]. Available: https://www.tandfonline.com/doi/full/10.1179/1476830515Y.0000000018

[43] P. K. Stein and Y. Pu, "Heart rate variability, sleep and sleep disorders," *Sleep Medicine Reviews*, vol. 16, no. 1, pp. 47–66, Feb. 2012. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1087079211000293

[44] R. Langner, F. Scharnowski, S. Ionta, C. E. G. Salmon, B. J. Piper, and G. S. P. Pamplona, "Evaluation of the reliability and validity of computerized tests of attention," *PLOS ONE*, vol. 18, no. 1, p. e0281196, Jan. 2023. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0281196

[45] S. T. Mueller and B. J. Piper, "The Psychology Experiment Building Language (PEBL) and PEBL Test Battery," *Journal of Neuroscience Methods*, vol. 222, pp. 250–259, Jan. 2014, arXiv: NIHMS150003 ISBN: 2122633255. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0165027013003762

[46] N. Narvaez Linares, V. Charron, A. Ouimet, P. Labelle, and H. Plamondon, "A systematic review of the Trier Social Stress Test methodology: Issues in promoting study comparison and replicable research," *Neurobiology of Stress*, vol. 13, p. 100235, Nov. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2352289520300254

[47] A. S. Perrotta, A. T. Jeklin, B. A. Hives, L. E. Meanwell, and D. E. Warburton, "Validity of the Elite HRV Smartphone Application for Examining Heart Rate Variability in a Field-Based Setting," *Journal of Strength and Conditioning Research*, vol. 31, no. 8, pp. 2296–2302, Aug. 2017, iSBN: 0000000000. [Online]. Available: http://journals.lww.com/00124278-201708000-00030

[48] Kubios, "User's Guide Kubios HRV Scientific," 2023.

[49] "Mindfield eSense Skin Response - GSR sensor for iPhone & Android." [Online]. Available: https://bio-medical.com/

mindfield-esense-skin-response-gsr-sensor-for-iphone-andriod.html

[50] H. Alshaher, "Studying the Effects of Feature Scaling in Machine Learning," Dissertation, North Carolina Agricultural and Technical State University, 2021.

[51] M. Radovic, M. Ghalwash, N. Filipovic, and Z. Obradovic, "Minimum redundancy maximum relevance feature selection approach for temporal gene expression data," *BMC Bioinformatics*, vol. 18, no. 1, p. 9, Dec. 2017. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1423-9

[52] G. Forte, F. Favieri, and M. Casagrande, "Heart rate variability and cognitive function: A systematic review," *Frontiers in Neuroscience*, vol. 13, no. JUL, pp. 1–11, 2019.

[53] C. Rudin and J. Radin, "Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition," *Harvard Data Science Review*, vol. 1, no. 2, Nov. 2019. [Online]. Available: https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/8

[54] M. Choi and J. J. Jeong, "Comparison of Selection Criteria for Model Selection of Support Vector Machine on Physiological Data with Inter-Subject Variance," *Applied Sciences (Switzerland)*, vol. 12, no. 3, 2022.

[55] W. Kraaij, S. Verberne, S. Koldijk, E. de Korte, S. van Dantzig, M. Sappelli, M. Shoaib, S. Bosems, R. Achterkamp, A. Bonomi, J. Schavemaker, B. Hulsebosch, T. Wabeke, M. Vollenbroek-Hutten, M. Neerincx, and M. v. Sinderen, "Personalized support for well-being at work: an overview of the SWELL project," *User Modeling and User-Adapted Interaction*, vol. 30, no. 3, pp. 413–446, 2020.

[56] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM, Aug. 2016, pp. 785–794.

[57] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015. [Online]. Available: https://www.nature.com/articles/nature14539

[58] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A Practical Guide to Support Vector Classification," Department of Computer Science National Taiwan University, Taipei, Taiwan, Tech. Rep., 2016.

[59] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0925231220307153

[60] B. Alam, "Naive Bayes Classifier Python Tutorial 2023," 2022. [Online]. Available: https://hands-on.cloud/naive-bayes-classifier-python-tutorial/

[61] L. Breiman, "Random Forest," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://link.springer.com/10.1023/A:1010933404324

[62] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *Lecture Notes in Computer Science*, Berlin, Heidelberg, 2000, vol. 1857, pp. 1–15, issue: 2 ISSN: 09509232. [Online]. Available: http://link.springer.com/10.1007/3-540-45014-9_1

[63] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. Sebastopol, California: O'Reilly Media, Inc, 2019.

[64] R. Castaldo, L. Montesinos, P. Melillo, C. James, and L. Pecchia, "Ultra-short term HRV features as surrogates of short term HRV: A case study on mental stress detection in real life," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, pp. 1–13, 2019, publisher: BMC Medical Informatics and Decision Making.