



Interpretable Machine Learning for Mental Workload Detection: From Binary to Multi-Class with Multimodal Physiology Signals

Ade Hendi¹, Auditya P. Sutarto^{2*}, Mega B. Herlambang³ and Nailul Izzah²

¹Department of Informatics Engineering, Universitas Qomaruddin, Gresik, Indonesia, 61152

²Department of Industrial Engineering, Universitas Qomaruddin, Gresik, Indonesia, 61152

³Department of Industrial Engineering, Institut Teknologi Indonesia, Tangerang Selatan, Indonesia, 15314

*Corresponding author: auditya@uqgresik.ac.id

Received 3 February 2026

Revised 28 April 2026

Accepted 8 May 2026

Abstract

Accurate detection of mental workload is essential for developing adaptive human–machine interaction (HMI) systems that respond to cognitive states of users in real time. Therefore, this study aimed to develop as well as evaluate machine learning models for classifying mental workload into binary, namely load vs. no load, and multi-class comprising low, medium, and high levels using multimodal physiological signals. Thirty-four participants completed two cognitively demanding tasks, and heart rate variability (HRV) as well as electrodermal activity (EDA) were recorded during the tasks. Signals were segmented into 1-min non-overlapping windows, with features extracted from the time, frequency, and nonlinear HRV domains, as well as tonic and phasic EDA indices. Moreover, three feature selection strategies, correlation-based, domain-expert, and minimum redundancy maximum relevance (mRMR), were evaluated across multiple classifiers, including gradient boosting, XGBoost, random forest, k-nearest neighbor, and multilayer perceptron. The model performance was evaluated using an 80/20 subject-wise train–test split, with 10-fold cross-validation conducted on the training set. Additionally, oversampling was applied exclusively to the training data to address the class imbalance. The result showed that multimodal fusion of HRV and EDA consistently outperformed single-modality models, achieving the strongest performance with XGBoost and correlation-based feature selection (binary: F1 = 0.787, AUC = 0.847; multi-class: Macro-F1 = 0.722, AUC = 0.862). The Shapley additive explanations (ShAP) analysis identified nonlinear HRV indices and phasic EDA features as the dominant predictors. These outcomes showed the importance of multimodal integration and data-driven feature selection for robust mental workload detection in HMI.

Keywords: electrodermal activity, heart rate variability, human-machine interaction, machine learning, mental workload detection, Shapley Additive explanations (ShAP)

1. Introduction

Human–machine interaction (HMI) is playing an increasingly central role in intelligent systems in healthcare, education, defense, and industry [1]. In the context of Industry 4.0 and 5.0, HMI surpasses efficiency improvement to ShAP human–machine collaboration in cognitively demanding and safety-critical environments. Therefore, ensuring system reliability requires the explicit consideration of human cognitive states, particularly mental workload, during system design and operation [1–3].

Based on the description above, the perspective prioritizes adaptive automation and instructional systems, dynamically adjusting task demands or support based on the state of the operator to prevent overload or under-stimulation [4], which depends on continuous as well as accurate workload monitoring. Therefore, real-time workload estimation is essential for adaptive systems that aim to optimize performance without compromising safety or well-being [2].

An excessive workload overwhelms attentional resources, weakens decision-making, and causes stress-like autonomic responses. Consequently, an insufficient workload leads to boredom, vigilance decrements, and

performance lapses [5]. These effects are particularly critical in safety-sensitive contexts, showing the need for reliable workload monitoring as a practical requirement for modern HMI or HCI systems.

Mental workload assessment has traditionally depended on subjective self-reports, such as the NASA Task Load Index (NASA-TLX), performance metrics, or a combination of both [3, 6]. These measures are retrospective, intrusive, and unsuitable for real-time adaptation. Performance-based indicators such as accuracy or response time provide task-specific perceptions but often confound workload with skill, strategy, or learning effects, limiting generalizability across contexts.

Physiological signals have appeared as a promising avenue for objective and continuous workload monitoring in response to the limitations mentioned above, including electrocardiograph-derived heart rate variability (HRV) and electrodermal activity (EDA/GSR) [4, 7–9]. These signals are attractive because of the sensitivity to autonomic nervous system dynamics, relative ease of acquisition, and compatibility with wearable devices [5]. Advances in sensor technology and machine learning have accelerated interest in physiological workload detection.

As shown in Table 1, previous studies on mental workload detection showed several consistent limitations in the methodological designs. Many studies depend on relatively long windows of 2–5 min [9–12], which provide stable physiological features but limit the temporal sensitivity and real-time applicability. More recent analyses have explored shorter windows ranging from 30 s to 1 min [13–15], enabling finer-grained workload tracking at the cost of increased signal variability.

Based on the description above, the experimental paradigms differ widely. Controlled laboratory tasks such as arithmetic, N-back, and task-switching are commonly used to stimulate workload under standardized conditions [11, 15, 16]. Meanwhile, stress-induction protocols such as the Trier Social Stress Test and Maastricht Acute Stress Test prioritize autonomic reactivity to stress [12, 17]. A smaller subset of studies has attempted to improve ecological validity through semi-realistic scenarios, including email writing and traffic control simulations [4, 18].

Feature selection practices range from early sequential and heuristic methods [13, 16] to more systematic data-driven strategies, such as minimum redundancy maximum relevance (mRMR), recursive feature elimination, and hybrid domain-informed strategies [11, 19, 20]. As many studies depend on standard k-fold cross-validation, more recent analyses adopt subject-wise or nested validation schemes to mitigate bias and data leakage [14, 17]. On the other hand, hyperparameter tuning remains inconsistently applied in relation to this discussion. Moreover, only a limited number of studies explicitly integrate shapley additive explanations (ShAP) or related methods, indicating that interpretability lags behind predictive performance as a study priority.

Previous analyses showed a lack of systematic and methodologically consistent evaluation of mental workload detection methods, particularly with window length selection, multimodal signal fusion, feature selection strategies, validation rigor, as well as model interpretability problems. Consequently, the process remains unclear, which combinations of signals and modeling selections provide robust, generalizable, as well as interpretable workload estimates that are suitable for real-time HMI applications.

The present study advances mental workload detection through experimental and methodological refinements to address the gaps. Compared to previous studies shown in Table 1, which predominantly focus on binary classification, single-modality signals, as well as limited interpretability, this study integrates multimodal physiological signals (HRV and EDA) with both binary and multi-class classification frameworks.

Table 1 Positioning of the Present Study Relative to Previous Analyses (2020–2025)

Aspect	Previous Studies	The present Study
Classification	Mostly binary [9, 9, 13–16, 18, 21]	Binary + multi-class
Signals	Single modality [4, 9, 14, 17, 18, 21]	HRV + GSR fusion
Window length	Often 2–5 min [9, 11, 12, 17]	Short window (1 min)
Validation	Often, a simple CV	Improved + clarified pipeline
Interpretability	Rare (few ShAP) [4, 11–13, 16, 18, 21]	ShAP integrated

The experimental protocol incorporates inter-task recovery periods to reduce carryover effects and adopts short, one-minute analysis windows to better support near-real-time monitoring. Methodologically, this study systematically evaluated binary and multi-class workload classification using HRV, EDA, and the multimodal fusion under three complementary feature selection strategies, namely correlation-based, mRMR, and domain-informed. The machine learning pipeline followed the best practices to ensure methodological rigor and model generalizability, including subject-wise validation, class imbalance handling, and leakage-resistant model evaluation [15]. Different from most previous studies, where interpretability remains limited, ShAP-based explanation is explicitly integrated to provide transparent and physiologically meaningful perceptions into feature contributions as well as model behavior.

2. Materials and Method

2.1 Dataset and Experimental Protocol

Thirty-four university students participated in the present study, including 17 males with a mean age = 21.2 ± 1.1 years and a mean BMI = 24.3 ± 5.8 . All individuals were right-handed and confirmed to maintain a regular sleep pattern the night before testing. Participants abstained from caffeinated drinks for at least two hours before the experiment to minimize potential confounders affecting autonomic function. In addition, participation was voluntary, and each participant signed a written informed consent form before data collection.

Participants completed two cognitive tasks designed to stimulate mental workload, which had been widely adopted in cognitive neuroscience and human factors studies to cause autonomic nervous system responses associated with mental workload [22, 23]. The first was a five-minute switcher task implemented using the Psychology Experiment Building Language (PEBL) software. In this task, participants identified pairs of figure elements according to a switching feature rule from 10 figures, each combining five colors, ShAPes, and letters shown on a black screen [22]. A brief practice session preceded the actual task, during which the recording was paused. The main task consisted of nine blocks, each requiring 12 correct responses before advancing. Both reaction time (RT) and accuracy were recorded for each trial, but were not used in this study.

The second task was the Iowa Gambling Task (IGT), a decision-making paradigm under uncertainty [24]. Participants were instructed to maximize the virtual earnings by selecting cards from four decks (A–D), each associated with a different pattern of reward and penalty. Decks A and B produced higher immediate gains but larger long-term losses, where decks C and D generated smaller short-term rewards and overall gains. Over 100 trials, participants received feedback after each selection and gradually adjusted individual strategy, engaged in risk-based learning processes, and increased the cognitive load.

Each experimental session consisted of five phases and lasted approximately 30–45 min, including Baseline → Task A (Switcher) → 5-min Rest → Task B (Iowa Gambling) → Recovery. Moreover, a 5-min rest interval was introduced to minimize carryover and accumulated workload, improving internal validity as well as task-specific attribution. All participants completed the experimental phases in the same order to maintain procedural consistency across participants, even though this design did not eliminate potential order effects.

Physiological data were collected during the analysis using two types of wearable sensors. Cardiovascular activity was recorded by applying a Polar H10 chest strap (Polar, Finland) at 130 Hz, and Electrocardiogram (ECG) signals were processed using the Elite HRV application to extract R–R intervals for HRV analysis (Perrotta et al., 2017). EDA was measured using an eSense EDA sensor (Mindfield Biosystem Ltd., Germany), with finger electrodes sampling at 5 Hz. Skin conductance was recorded in microsiemens (μS) and exported for offline processing. During the process, the signal quality was checked before each session.

Following each task, participants completed the NASA-TLX, a widely used multidimensional measure of perceived workload. The questionnaire assessed six subscales, namely, mental demand, physical demand, temporal demand, performance, effort, and frustration, which were rated on a 0–100 scale. In this study, the NASA-TLX scores served as the ground truth for labeling workload levels.

The baseline, rest, and recovery phases during which no cognitive task was performed were labelled as low workload. For the two task conditions, workload levels were assigned using an in-subject ranking method. For each participant, the task with the higher NASA-TLX score was labelled as high workload, while the other was labelled as medium. This method avoided reliance on arbitrary global thresholds and considered interindividual differences in subjective perception. Physiological signals were segmented into 1-min non-overlapping windows, and each segment inherited the workload label of the corresponding task phase. Consequently, all segments in a given task block were assigned the same workload.

Data acquisition was time-synchronized on a laptop with markers indicating the start and end of each experimental phase. The participants were seated in a standardized posture under controlled environmental conditions, namely, time of day, noise, temperature, and lighting. After receiving standardized instructions and sensor placement, the participants completed the task sequence, with the NASA-TLX administered immediately after each task block.

2.2 Signal Processing and Feature Extraction

ECG signals were captured using the Elite HRV application and exported to the Kubios HRV software for further analysis [25]. Following established HRV preprocessing guidelines [26], artifact correction was applied using a medium filter setting to remove ectopic beats and correct irregular intervals through interpolation, ensuring physiologically reliable RR interval series before feature extraction. Similarly, the EDA signals were preprocessed using the built-in filtering procedures of the device before EDA feature extraction. The signals were then segmented into non-overlapping 1-min windows to increase the temporal resolution. This segmentation led to

multiple samples per participant, as the effective dataset size reflected window-level observations rather than participant-level counts.

Kubios extracted HRV indices across three domains for each segment, namely, time-domain (meanNN, SDNN, RMSSD, pNN50), frequency-domain (low frequency [LF], high frequency [HF], and LF/HF ratio), as well as nonlinear measures (SD1, SD2). These indices captured the autonomic balance and complexity of cardiovascular dynamics. Table 2 showed a description of the EDA and HRV features during this study.

EDA was recorded using the eSense GSR sensor, with raw skin conductance data exported in microsiemens for offline processing of the data. The signals were decomposed into tonic and phasic components using the NeuroKit2 Python library [27]. Similar to HRV, EDA signals were segmented into non-overlapping 1-min windows. For each segment, standardized and device-agnostic features were computed, including tonic level statistics (mean and SD), phasic activity metrics (number of skin conductance responses [SCRs], mean SCR amplitude, and amplitude variability), and additional variability measures.

Table 2 Description of EDA and HRV features

Parameter	Description
Galvanic Skin Response	
scr_peaks (count)	Number of skin conductance responses (SCRs) detected in the time window.
scr_peaks_amp_mean (μ S)	The mean amplitude of the detected SCRs reflected the response intensity.
eda_tonic_sd (μ S)	Standard deviation of the tonic (baseline) activity, indicating arousal variability.
eda_phasic_sd (μ S)	Standard deviation of the phasic signal (event-related).
eda_amp_mean (μ S)	Mean of the total (raw) EDA signal, including the tonic and phasic components.
eda_amp_sd (μ S)	Standard deviation of the raw EDA signal
Heart Rate Variability	
LF power (ms^2)	Absolute power of the low-frequency band (0.04–0.15 Hz)
LF power (ms^2)	Relative power of the low-frequency band (0.04–0.15 Hz) in normalized units
HF power (ms^2)	Absolute power of the high-frequency band (0.15–0.4 Hz)
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
Mean RR (ms)	Mean of all RR intervals (HR x RR interval = 60,000)
SDNN (ms)	Standard deviation of NN intervals
pNN50 (%)	Percentage of successive RR-intervals differing by more than 50 ms
RMSSD	Root mean square of successive NN interval differences
SD1 (ms)	The width of the Poincaré plot
SD2 (ms)	The length of the Poincaré plot

2.3 Data pre-processing

The dataset was split into training (80%) and test (20%) sets using stratified subject-wise sampling to ensure that no participant appeared in both sets, preventing data leakage. The held-out test set was reserved exclusively for the final model evaluation and was not used during the model development. For binary classification, the dataset included 734 no-load (56.5%) and 566 load (43.5%) instances. For the multiclass setting, the dataset comprised 734 low- (56.5%), 286 medium- (22.0%), and 280 high-load (21.5%) instances, indicating a moderate class imbalance, particularly for medium- and high-workload states.

The Adaptive Synthetic Sampling (ADASYN) algorithm was applied to the training set only to address the problem, generating synthetic samples for minority classes while preserving the original distribution of the test set. Compared to conventional SMOTE, ADASYN prioritized harder-to-learn minority samples, reducing the bias toward easily classified cases and improving the robustness of the model under imbalanced conditions [28].

Three complementary feature selection strategies were evaluated as preprocessing steps applied to the training data only in the inner cross-validation loop. The optimal configuration identified during the inner CV was refitted to the full training set and evaluated on the held-out test set. The feature selection strategies used during the process included the following. First, a correlation-based filter that removed highly correlated features ($|r| > 0.8$) to reduce redundancy. Second, minimum redundancy maximum relevance (mRMR), which selected features maximizing relevance to class labels while minimizing inter-feature dependence [29]. Third, a domain-informed subset based on physiologically relevant HRV and EDA features reported in a previous workload study [26, 30].

Each strategy was applied independently to the HRV-only, EDA-only, and fused HRV+EDA feature sets, ensuring systematic feature pruning across the unimodal as well as multimodal conditions. This design enabled a transparent comparison of data-driven and domain-based feature selection methods and mitigated the overfitting risks associated with feature fusion. All preprocessing steps, including min-max scaling and one-hot encoding, were embedded in the inner cross-validation pipeline to prevent data leakage. Moreover, data were collected under controlled laboratory conditions, and signal preprocessing (artifact correction and segmentation) ensured that all retained data windows contained valid physiological features. During this process, no samples were excluded because of missing values.

2.4 Machine Learning Pipeline

Figure 2 showed the complete machine learning pipeline implemented in this study, which spanned raw signal acquisition, labeling, feature extraction, class balancing, feature selection, preprocessing, model training with nested cross-validation, performance evaluation, and ShAP-based explanation. The analysis implemented four machine learning classifiers representing probabilistic, discriminative, ensemble, and neural methods, selected based on the widespread as well as well-established use in machine learning-based cognitive workload detection studies. Further details on the applications and effectiveness in physiological signal analysis could be found in a previous comprehensive review [31].

1. Support Vector Machine (SVM). SVM is a discriminative model that aims to find the optimal hyperplane separating data points into distinct classes and is particularly effective in high-dimensional spaces.
2. K-Nearest Neighbors (KNN). KNN is a non-parametric, instance-based learning algorithm that classifies new samples based on the majority label of the nearest neighbors in the feature space. Distance metrics, such as Euclidean distance, determine neighborhood similarity, and the number of neighbors k acts as a smoothing parameter.
3. Extreme Gradient Boosting (XGBoost). XGBoost is an advanced ensemble-based boosting method that extends gradient boosting with regularization and system optimization. It builds decision trees sequentially, where each new tree corrects the errors of the previous ones but adds penalization terms to prevent overfitting.
4. Multilayer Perceptron (MLP). The MLP is a feedforward artificial neural network composed of an input layer, one or more hidden layers, and an output layer. Neurons in the hidden layers apply nonlinear activation functions to weighted inputs, allowing the model to capture complex relationships in the data set.

All classifiers were implemented in Python (v3.10) using the scikit-learn library (v1.4), except for the MLP, which was developed using TensorFlow (v2.15)/Keras for flexible neural network training. GridSearchCV from Scikit-learn handled hyperparameter tuning for all models embedded in a nested cross-validation framework.

Model hyperparameters were optimized exclusively on the training data to prevent information leakage. Tuning was performed using GridSearchCV with a 10-fold stratified K-fold procedure embedded in a subject-wise nested cross-validation framework. Table 3 showed the optimal hyperparameters identified for each classifier in both binary and multi-class settings. The selected configurations reflected a balance between model complexity and generalization, as informed by previous studies [9, 32], empirical evaluation during cross-validation, and established best practices in physiological signal-based workload detection. After selecting the optimal configuration, the model was refitted to the full training data and evaluated once on the held-out test set. Following this process, binary classification was evaluated using Accuracy, F1-score, and AUC-ROC. The macro-averaged F1 and AUC-ROC were reported to ensure a balanced evaluation across workload levels for multi-class classification.

The model performance variability was assessed using confidence intervals derived from the cross-validation results in addition to the point estimates. Statistical comparisons between the models were conducted using the Wilcoxon signed-rank test across folds to evaluate the significance of the observed differences.

This study applied shapley additive explanations (ShAP) to the final fitted models across all conditions to improve model interpretability. Grounded in cooperative game theory, ShAP quantified the positive or negative contribution of each feature to the model predictions [33]. Global bar and review plots were used to rank influential HRV and EDA features. In addition, beeswarm plots showed how feature values affected class probabilities, and local ShAP values explained the individual predictions. This multilevel analysis provided both population-level perceptions and subject-specific explanations, extending the interpretability outside the conventional feature importance measures. In the context of physiological workload detection, ShAP enabled a direct connection between model outputs and interpretable physiological markers, such as identifying how RMSSD or skin conductance consistently contributed to higher workload predictions.

Table 3 Optimal hyperparameters for the best-performing classifiers

Classifier	Binary	Multi
svm	$C = 1$, kernel = poly, degree = 5, gamma = 1	$C = 0.1$, kernel = linear, degree = 1, gamma = 1
knn	n_neighbors = 4, weights = distance, algorithm = auto	n_neighbors = 4, weights = uniform, algorithm = auto
xgb	learning_rate = 0.1, max_depth = 7, gamma = 0.2	learning_rate = 0.05, max_depth = 5, gamma = 0
mlp	activation = ReLU, hidden_layer_sizes = (400,), solver = adam	activation = ReLU, hidden_layer_sizes = (100,), solver = sgd

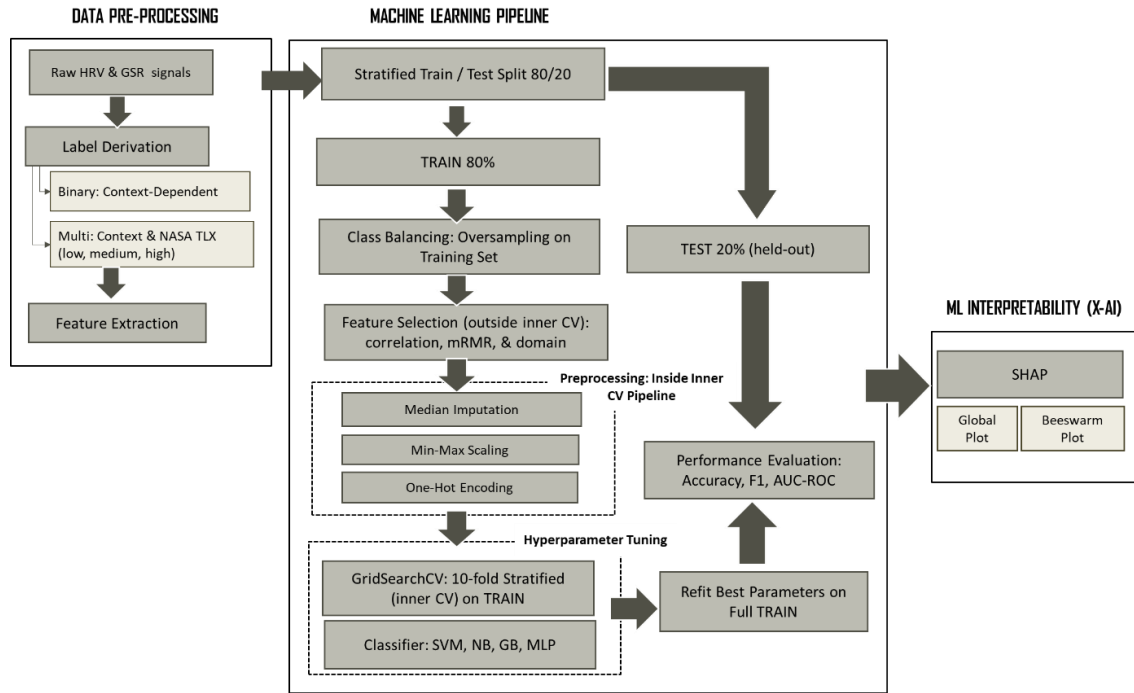


Figure 1 Overview of the machine learning pipeline for mental workload detection

3. Results and Discussion

Table 4 showed that across both tasks, the fusion of HRV and EDA consistently outperformed the single modalities, achieving the strongest results with XGBoost and correlation-based feature selection (Binary: F1 = 0.787, AUC = 0.847; multi-class: F1 = 0.722, AUC = 0.862). HRV proved more effective than EDA, particularly in binary classification, where Gradient Boosting reached F1 = 0.751 and AUC = 0.800. EDA models were weaker in binary but competitive in multi-class classification (Macro-F1 = 0.722, AUC = 0.807). Among the feature selection strategies, correlation-based selection consistently produced the best results, outperforming both the domain expert and mRMR methods. These results showed the value of multimodal fusion and data-driven feature selection in developing reliable and generalizable mental workload detection models.

Table 4 Review of Classification Results having mean values with 95% confidence intervals in parentheses

Feature Select	Model	Acc	Binary F1	AUC	Acc	Multi F1	AUC
HRV							
Correlation	svm	0.682 (0.620, 0.744)	0.696 (0.624, 0.768)	0.746 (0.682, 0.810)	0.456 (0.379, 0.533)	0.443 (0.366, 0.520)	0.636 (0.566, 0.706)
	knn	0.668 (0.603, 0.733)	0.689 (0.608, 0.770)	0.698 (0.633, 0.763)	0.686 (0.619, 0.753)	0.682 (0.615, 0.749)	0.765 (0.714, 0.816)
	xgb	0.731 (0.661, 0.801)	0.723 (0.637, 0.809)	0.800 (0.733, 0.867)	0.674 (0.611, 0.737)	0.674 (0.611, 0.737)	0.840 (0.780, 0.900)
	mlp	0.671 (0.601, 0.741)	0.689 (0.604, 0.774)	0.740 (0.675, 0.805)	0.466 (0.396, 0.536)	0.465 (0.395, 0.535)	0.641 (0.585, 0.697)
	svm	0.608 (0.550, 0.666)	0.674 (0.593, 0.755)	0.669 (0.598, 0.740)	0.440 (0.367, 0.513)	0.418 (0.345, 0.491)	0.631 (0.57, 0.692)
mRMR	knn	0.675 (0.616, 0.734)	0.687 (0.626, 0.748)	0.750 (0.677, 0.823)	0.606 (0.549, 0.663)	0.602 (0.545, 0.659)	0.704 (0.655, 0.753)
	xgb	0.724 (0.669, 0.779)	0.727 (0.648, 0.806)	0.793 (0.736, 0.850)	0.514 (0.433, 0.595)	0.510 (0.429, 0.591)	0.706 (0.646, 0.766)
	mlp	0.654 (0.592, 0.716)	0.686 (0.615, 0.757)	0.693 (0.623, 0.763)	0.475 (0.408, 0.542)	0.473 (0.406, 0.540)	0.635 (0.574, 0.696)
	svm	0.579 (0.510, 0.648)	0.582 (0.458, 0.706)	0.626 (0.551, 0.701)	0.400 (0.326, 0.474)	0.395 (0.321, 0.469)	0.559 (0.497, 0.621)
	knn	0.560 (0.486, 0.634)	0.530 (0.452, 0.608)	0.649 (0.572, 0.726)	0.667 (0.609, 0.725)	0.662 (0.604, 0.720)	0.752 (0.704, 0.800)
Domain	xgb	0.635 (0.571, 0.699)	0.604 (0.535, 0.673)	0.679 (0.612, 0.746)	0.547 (0.481, 0.613)	0.547 (0.481, 0.613)	0.742 (0.696, 0.788)
	mlp	0.624 (0.563, 0.685)	0.630 (0.552, 0.708)	0.672 (0.606, 0.738)	0.409 (0.328, 0.490)	0.404 (0.323, 0.485)	0.572 (0.522, 0.622)
EDA							
Correlation	svm	0.591 (0.543, 0.639)	0.387 (0.266, 0.508)	0.522 (0.210, 0.834)	0.328 (0.186, 0.470)	0.211 (0.069, 0.353)	0.454 (0.394, 0.514)
	knn	0.584 (0.551, 0.677)	0.524 (0.456, 0.592)	0.626 (0.546, 0.706)	0.601 (0.546, 0.656)	0.600 (0.545, 0.655)	0.701 (0.641, 0.761)
	xgb	0.577 (0.522, 0.632)	0.527 (0.459, 0.595)	0.628 (0.545, 0.711)	0.537 (0.474, 0.600)	0.537 (0.474, 0.600)	0.716 (0.671, 0.761)
	mlp	0.581 (0.516, 0.646)	0.391 (0.319, 0.463)	0.558 (0.460, 0.656)	0.287 (0.186, 0.388)	0.194 (0.093, 0.295)	0.493 (0.425, 0.561)
	svm	0.583 (0.535, 0.631)	0.367 (0.246, 0.488)	0.538 (0.226, 0.850)	0.353 (0.211, 0.495)	0.255 (0.113, 0.397)	0.486 (0.426, 0.546)
mRMR	knn	0.577 (0.514, 0.640)	0.484 (0.416, 0.552)	0.615 (0.535, 0.695)	0.693 (0.635, 0.751)	0.689 (0.631, 0.747)	0.770 (0.714, 0.826)
	xgb	0.633 (0.582, 0.684)	0.578 (0.520, 0.636)	0.656 (0.579, 0.733)	0.560 (0.501, 0.619)	0.558 (0.499, 0.617)	0.759 (0.715, 0.803)
	mlp	0.584 (0.522, 0.646)	0.442 (0.369, 0.515)	0.561 (0.447, 0.675)	0.358 (0.275, 0.441)	0.290 (0.207, 0.373)	0.551 (0.481, 0.621)

Note. The highest AUC and F1 scores within each modality are highlighted in **bold**, and the top results within each feature selection method are shown in *italics*.

Table 4 Review of Classification Results having mean values with 95% confidence intervals in parentheses (Cont.)

Feature Select	Model	Acc	Binary F1	AUC	Acc	Multi F1	AUC
EDA							
Domain	svm	0.549 (0.514, 0.584)	0.189 (0.068, 0.310)	0.443 (0.160, 0.726)	0.406 (0.328, 0.484)	0.395 (0.317, 0.473)	0.586 (0.522, 0.650)
	knn	0.466 (0.409, 0.523)	0.366 (0.297, 0.435)	0.465 (0.390, 0.540)	0.726 (0.650, 0.802)	0.722 (0.649, 0.795)	0.795 (0.742, 0.848)
	xgb	0.553 (0.667, 0.799)	0.431 (0.669, 0.797)	0.547 (0.666, 0.866)	0.629 (0.556, 0.702)	0.629 (0.553, 0.705)	0.807 (0.746, 0.868)
	mlp	0.541 (0.599, 0.711)	0.378 (0.558, 0.716)	0.539 (0.574, 0.838)	0.394 (0.322, 0.466)	0.388 (0.316, 0.460)	0.595 (0.541, 0.649)
HRV + EDA							
Correlation	svm	0.702 (0.646, 0.758)	0.696 (0.632, 0.760)	0.760 (0.699, 0.821)	0.452 (0.376, 0.528)	0.441 (0.365, 0.517)	0.642 (0.577, 0.707)
	knn	0.670 (0.611, 0.729)	0.689 (0.623, 0.755)	0.731 (0.670, 0.792)	0.727 (0.659, 0.795)	0.722 (0.654, 0.790)	0.796 (0.736, 0.856)
	xgb	0.787 (0.728, 0.846)	0.787 (0.727, 0.847)	0.847 (0.782, 0.912)	0.713 (0.643, 0.783)	0.713 (0.643, 0.783)	0.862 (0.804, 0.920)
	mlp	0.748 (0.681, 0.815)	0.753 (0.672, 0.834)	0.801 (0.735, 0.867)	0.472 (0.409, 0.535)	0.472 (0.409, 0.535)	0.641 (0.588, 0.694)
mRMR	svm	0.617 (0.563, 0.671)	0.671 (0.614, 0.728)	0.667 (0.595, 0.739)	0.433 (0.364, 0.502)	0.406 (0.337, 0.475)	0.627 (0.566, 0.688)
	knn	0.709 (0.644, 0.774)	0.711 (0.647, 0.775)	0.787 (0.722, 0.852)	0.656 (0.593, 0.719)	0.651 (0.588, 0.714)	0.742 (0.690, 0.794)
	xgb	0.691 (0.625, 0.757)	0.715 (0.645, 0.785)	0.743 (0.675, 0.811)	0.665 (0.603, 0.727)	0.665 (0.603, 0.727)	0.823 (0.768, 0.878)
	mlp	0.642 (0.579, 0.705)	0.683 (0.611, 0.755)	0.695 (0.626, 0.764)	0.470 (0.387, 0.553)	0.467 (0.384, 0.550)	0.633 (0.568, 0.698)
Domain	svm	0.569 (0.500, 0.638)	0.598 (0.530, 0.666)	0.613 (0.540, 0.686)	0.406 (0.328, 0.484)	0.395 (0.317, 0.473)	0.586 (0.522, 0.650)
	knn	0.648 (0.575, 0.721)	0.650 (0.582, 0.718)	0.689 (0.621, 0.757)	0.726 (0.653, 0.799)	0.722 (0.649, 0.795)	0.795 (0.742, 0.848)
	xgb	0.733 (0.672, 0.794)	0.733 (0.668, 0.798)	0.766 (0.701, 0.831)	0.629 (0.553, 0.705)	0.629 (0.553, 0.705)	0.807 (0.746, 0.868)
	mlp	0.655 (0.587, 0.723)	0.637 (0.566, 0.708)	0.706 (0.639, 0.773)	0.394 (0.322, 0.466)	0.388 (0.316, 0.460)	0.595 (0.541, 0.649)

Note. The highest AUC and F1 scores within each modality are highlighted in **bold**, and the top results within each feature selection method are shown in *italics*.

The results supported previous studies indicating that combining physiological modalities improved discriminative performance in mental workload detection [11, 14, 16, 17]. At the same time, HRV achieved acceptable results across both binary (AUC 0.69–0.80) and multi-class settings (AUC 0.63–0.84), confirming earlier outcomes that cardiac variability served as a reliable single-source predictor of cognitive and mental workload. Consequently, models depending solely on EDA produced weaker performance, consistent with previous evidence that EDA was less stable for classification [16, 34].

As the multiclass models produced higher AUC values than the binary classifiers, the accuracy, precision, recall, and F1 scores were comparatively lower. This apparent discrepancy arose because the AUC evaluated the quality of probability ranking rather than the accuracy of hard class assignments. In multiclass settings, where the AUC was often computed as the mean of one-vs-rest or pairwise comparisons, the models achieved good probability separation across classes, leading to inflated AUC values. After these probabilities were collapsed into single predictions, misclassifications became more possible, particularly considering the class imbalance and the inherently lower chance level in three-class problems. The models showed strong discriminatory ability in terms of ranking probabilities but weaker performance when considering strict classification outcomes.

An important practical consideration was that distinguishing medium from high workload levels often presented greater difficulty than distinguishing low from high workload levels. This reflected the reality of overlapping physiological responses at adjacent workload intensities, where variability in both HRV and EDA could obscure class boundaries. The results showed that the fusion of HRV and EDA provided the greatest benefit in these contexts. Moreover, the complementary signals resolved finer distinctions that single modalities might miss. The outcome showed the practical utility of multimodal systems for applications in which monitoring graded workload states, rather than simple load versus no load, was essential.

Confidence intervals were computed to quantify the performance variability beside the point estimates, as shown in Table 4. Across the models, the relatively narrow intervals indicated stable performance during the process. Pairwise comparisons using the Wilcoxon signed-rank test indicated that XGBoost achieved the highest mean performance and was significantly superior to all other classifiers ($p < 0.05$). Moreover, the results indicated that shorter analysis windows (1 min) produced better model performance than the 5-min windows commonly reported in earlier studies (e.g., [10, 18, 21]). This reflected greater sensitivity to rapid workload fluctuations, as mental workload varied more dynamically than slower evolving states, such as fatigue. The superior performance of shorter windows might also partly reflect reduced carryover between conditions and clearer signal attribution, particularly because rest periods were embedded in the experimental protocol. Shorter windows were less expected to capture residual arousal from the preceding tasks by minimizing the temporal span of each analysis segment, isolating workload-related responses more effectively. As this interpretation was plausible, the present study recognized that the observed gains could also be influenced by statistical inflation due to the larger number of patients. Future studies should determine how this success reflects true cognitive sensitivity or statistical inflation from having more data points.

3.1 ShAP-based Feature Interpretability

Figure 2 showed that for the binary model distinguishing workload from non-workload, the ShAP values indicated both HRV and EDA features as influential. The most impactful predictors were nonlinear HRV indices, namely the SD2/SD1 ratio and SD2, which captured the complexity as well as long-term variability of heart rate dynamics. These were followed by EDA-related features such as the mean amplitude of skin conductance responses (`eda_amp_men`), number of SCR peaks, and amplitude variability (`eda_amp_sd`), alongside HRV metrics including SD1 and maximum heart rate. The beeswarm plot showed the consistent positive contributions of higher SCR peaks and elevated HRV complexity (high SD2/SD1) in predicting workload. This outcome indicated that both autonomic cardiac variability and phasic electrodermal responses were reliable discriminators in the binary setting.

When extending to low, medium, and high workload classes, the ShAP ranking shifted slightly, even though the dominance of nonlinear HRV indices remained clear. SD2/SD1 and SD2 appeared as the strongest predictors, but the importance of frequency-domain HRV features became more pronounced, particularly peak VLF and HF_nu. Electrodermal indices, such as SCR peaks and `eda_amp_men`, also contributed significantly, as the effects appeared more heterogeneous across workload levels. The beeswarm distribution showed that high values of SCR peaks and `eda_amp_men` tended to drive predictions toward higher workload classes, while lower variability features, such as low HF_nu, shifted outputs toward reduced workload.

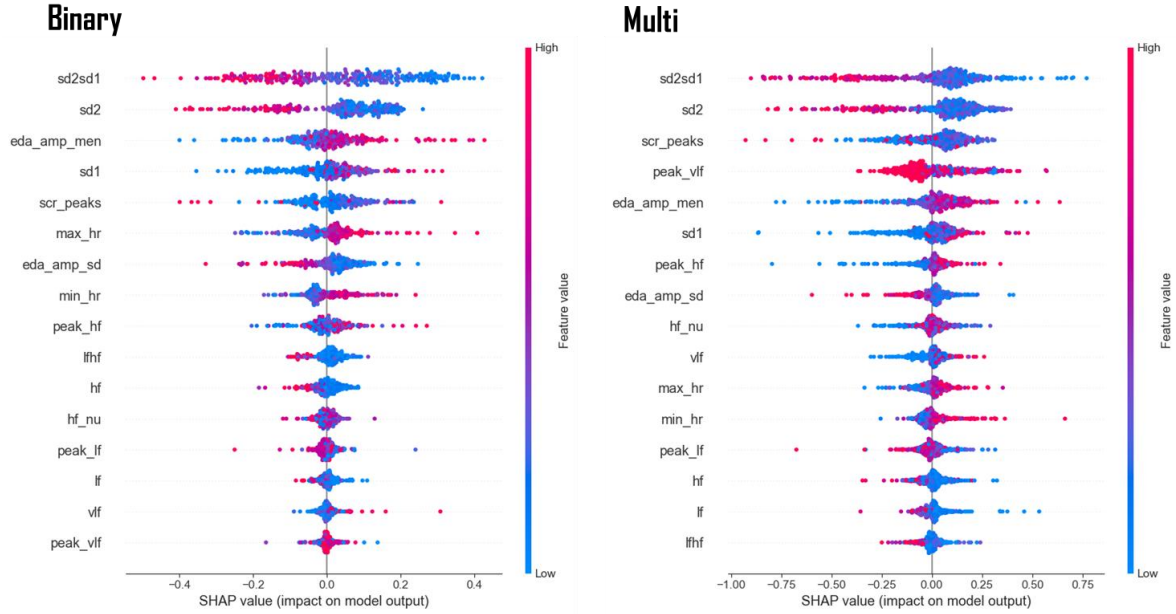


Figure 2 ShAP value distribution (beeswarm) plot showing feature contributions for binary (left) and multi-class (right) workloads for the best classification models. Each point represented an individual observation, with the color indicating the feature value from low to high.

Both models consistently identified nonlinear HRV indices (SD2/SD1, SD2, and SD1) as the most influential features, showing the role of heart rate complexity in reflecting mental workload. For HRV, Poincaré plot-derived indices, particularly the SD2/SD1 ratio, reflected a fundamental shift in cardiac autonomic regulation under increasing cognitive demand [35]. At rest, short-term beat-to-beat variability (SD1) was driven predominantly by parasympathetic/vagal modulation. Meanwhile, SD2 captured longer-term variability, reflecting combined sympathovagal contributions, showing that a low SD2/SD1 ratio indicated intact vagal dominance. Under mental workload, the prefrontal cortex applied an inhibitory influence on vagal outflow through the central autonomic network (CAN), disproportionately suppressing SD1 relative to SD2 [26]. These indices captured the dynamic structure of autonomic regulation beyond the representation of linear measures. The prominence in the ShAP rankings showed that mental workload reduced autonomic complexity and pushed the cardiovascular system toward more constrained, sympathetically dominated regulation, reflecting diminished adaptability under cognitive strain.

The most important EDA features were primarily amplitude-based measures and SCR peaks, which reflected discrete sympathetic cholinergic activation of eccrine sweat glands driven by descending hypothalamic pathways [30]. Compared to HRV-derived indices, which capture parasympathetic withdrawal, the SCR amplitude guided the intensity of individual sympathetic activation bursts, as the peak frequency reflected the rate of autonomic recruitment under load. In binary classification, the influence of phasic EDA activity, including SCR peaks and amplitude measures, was stronger, indicating that short-term autonomic arousal was a major signal for distinguishing the workload from the baseline states. Consequently, the multi-class model integrated a broader range of features, with frequency-domain HRV components (VLF, HF_nu) becoming more important, reflecting the ability to capture graded distinctions across low, medium, and high workloads. This comparison showed that binary models prioritized arousal-linked signals (EDA peaks), and multiclass models depended more on HRV complexity as well as frequency content to differentiate finer workload levels.

The ShAP results showed two complementary processes arising from partially independent physiological pathways, cardiac autonomic modulation and electrodermal sympathetic activation, that jointly reflected the response of the organism to cognitive demand. This complementarity explained the reason the multimodal HRV–EDA model consistently outperformed unimodal methods, as each modality contributed distinct, non-redundant information about workload-related autonomic dynamics.

This study had several implications for the design of workload detection systems. In small-sample HMI studies, shorter analysis windows provided practical advantages by capturing rapid workload fluctuations and increasing the number of training instances, even though the investigations might introduce a sample dependence. Longer windows were more suitable for slower-evolving states, such as fatigue, or for large datasets, as the features improved signal stability and reduced noise. Moreover, the window length needed to follow the temporal dynamics of the target cognitive state, with shorter windows supporting real-time monitoring and longer windows capturing gradual processes.

The multimodal fusion of HRV and EDA consistently outperformed the unimodal methods beyond temporal considerations, showing the complementary information transferred by the cardiac and electrodermal responses. As HRV remained a robust core modality, standardized EDA features added valuable arousal-related information that improved discrimination, particularly in the binary classification. In addition, correlation-based feature selection showed superior generalization compared with the domain-driven and mRMR methods, indicating that data-driven selection improved performance without sacrificing interpretability. The use of standardized physiological indices combined with ShAP-based explanations further strengthened the transparency and cross-study comparability. Collectively, these results supported the development of multimodal, data-driven, and interpretable workload monitoring systems that could be adapted to diverse HMI contexts.

3.2 Limitations and Future Directions

This study had several limitations during the process of the analysis. The relatively small and homogeneous sample restricted the generalizability of the findings across populations and tasks. In addition, data collection was conducted in a laboratory setting, which did not fully capture the complexity and variability of real-world task performances. A further limitation was the same task sequence, which introduced learning, adaptation, or fatigue effects despite the inclusion of a 5-min rest period. Therefore, distinctions between medium and high workload should be interpreted with caution. Building on these limitations, future studies should extend the analysis toward ecologically valid and passive monitoring scenarios, including real-world deployments, where latency, computational constraints, as well as user acceptance of wearable sensors are critical. Larger and more diverse cohorts are required to improve robustness and cross-population generalizability. Expanding the signal suite to include additional modalities, namely pupillometry, EEG, or facial expression analysis, may further strengthen multimodal inference as well as address domain-shift challenges across time and settings. Finally, the 1-min analysis interval should be validated in real-world contexts to ensure an appropriate balance between responsiveness and computational load.

Future study should explore the integration of interpretable workload detection models into adaptive HMI systems at a broader level, where real-time feedback can dynamically adjust task demands, interface complexity, and operator support. From a methodological perspective, hybrid methods that combine data-driven machine learning with physiologically informed models need investigation to improve robustness when applied to different users, tasks, or environments than those on which the methods are trained. Finally, systematically evaluating the trade-offs between interpretability, computational efficiency, and predictive performance is critical for translating these models into deployable real-world systems.

4. Conclusion

In conclusion, distinguishing medium from high workload remained challenging due to overlapping physiological responses. On the other hand, the multimodal fusion of HRV and EDA consistently improved discrimination. The strongest performance was achieved by multimodal models, particularly when combined with correlation-based feature selection and ensemble classifiers such as XGBoost. HRV provided a robust baseline, while EDA contributed most effectively when integrated with HRV, improving the predictive performance and interpretability. The use of standardized EDA indices further supported cross-study comparability.

This study contributed to the field by systematically evaluating binary and multi-class workload classification using multimodal physiological signals in an interpretable machine learning framework supported by a rigorously designed evaluation pipeline. From a practical perspective, the results showed that short analysis windows could support near-real-time workload monitoring in small-sample HMI contexts, even though the processes require careful consideration of the signal variability. The integration of multimodal sensing with ShAP-based interpretability enabled transparent and actionable perceptions, supporting the development of adaptive HMI systems in cognitively demanding environments.

5. Ethical Approval

This study followed the principles of the Declaration of Helsinki (1964) and received ethical clearance from the Institutional Health Study Ethics Committee (3916/KEP-UNISA/VIII/2025).

6. Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this study, the authors used ChatGPT 5.2 to refine sentences for better clarity, tone, and grammar, and to organize the literature review. After using the tool/service, the authors reviewed and edited the content as needed and were fully responsible for the content of the publication.

7. Acknowledgements

The authors are grateful to the Directorate of Indonesian Higher Education under the Study Grant for Junior Lectures [Hibah Penelitian Dosen Pemula No: 128/C3/DT.05.00/PL/2025] for the support.

8. Author Contributions

Hendi, A: Conceptualization, Methodology, Funding acquisition, Writing – review & editing, Supervision; Sutarto, AP: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft; Herlambang MB: Methodology, Formal analysis, Software, Validation, Writing – review & editing; Izzah, N: Conceptualization, Investigation, Project administration, Resources, Writing – review & editing.

9. References

- [1] Iarlori S, Perpetuini D, Tritto M, et al. An Overview of Methods and Methods for the Cognitive Workload Estimation in Human–Machine Interaction Scenarios through Wearables Sensors. *BioMedInformatics*. 2024;4:1155–1173.
- [2] Ramakrishnan P, Balasingam B, Biondi F. Cognitive load estimation for adaptive human–machine system automation. In: *Learning Control*. Elsevier; 2021. p. 35–58.
- [3] Longo L, Wickens CD, Hancock G, et al. Human Mental Workload: A Survey and a Novel Inclusive Definition. *Front Psychol*. 2022;13:883321.
- [4] Hillege RHL, Lo JC, Janssen CP, et al. The Mental Machine: Classifying Mental Workload State from Unobtrusive Heart Rate-Measures Using Machine Learning. In: Sottolare RA, Schwarz J (eds) *Adaptive Instructional Systems*. Cham: Springer International Publishing; 2020. p. 330–349.
- [5] Wang P, Houghton R, Majumdar A. Detecting and Predicting Pilot Mental Workload Using Heart Rate Variability: A Systematic Review. *Sensors*. 2024;24:3723.
- [6] Tierney M, Adeyemi EO, Bommer S. A Systematic Literature Review: Cognitive Workload Assessment in Human Factors Study. *Adv Hum-Comput Interact*. 2025;2025:9313239.
- [7] Dalmeida KM, Masala GL. HRV Features as Viable Physiological Markers for Stress Detection Using Wearable Devices. *Sensors*. 2021;21:2873.
- [8] Sardar SK, Kumar N, Lee SC. A Systematic Literature Review on Machine Learning Algorithms for Human Status Detection. *IEEE Access*. 2022;10:74366–74382.
- [9] Sutarto AP, Bagus Herlambang M, Izzah N, et al. Optimizing Mental Workload Detection for HCI: Comparative Feature Selection and Interpretable Machine Learning. *Int J Comput Digit Syst*. 2025;17:1–17.
- [10] Gjoreski M, Luštrek M, Gams M, et al. Monitoring stress with a wrist device using context. *J Biomed Inform*. 2017;73:159–170.
- [11] Izzah N, Sutarto AP, Hendi A, et al. Physiological Signals as Predictors of Mental Workload: Evaluating Single Classifier and Ensemble Learning Models. *J Optimasi Sist Ind*. 2023;22:81–98.
- [12] Pettersson K, Tervonen J, Narvainen J, et al. Selecting Feature Sets and Comparing Classification Methods for Cognitive State Estimation. In: *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*; 2020. p. 683–690.
- [13] Ahmed MU, Begum S, Gestlöf R, et al. Machine Learning for Cognitive Load Classification – A Case Study on Contact-Free Method. In: Maglogiannis I, Iliadis L, Pimenidis E (eds) *Artificial Intelligence Applications and Innovations*. Cham: Springer International Publishing; 2020. p. 31–42.
- [14] Bahameish M, Stockman T, Requena Carrión J. Strategies for Reliable Stress Recognition: A Machine Learning Method Using Heart Rate Variability Features. *Sensors*. 2024;24:3210.
- [15] Dayal A, Khan Z, Ullah K, et al. A Novel Method to Cognitive Load Measurement in N-Back Tasks Using Wearable Sensors, Empirical Mode Decomposition With Machine Learning, and Explainable AI for Feature Importance. *IEEE Access*. 2025;13:17083–17098.
- [16] Ding Y, Cao Y, Duffy VG, et al. Measurement and identification of mental workload during simulated computer tasks with multimodal methods and machine learning. *Ergonomics*. 2020;63:896–908.
- [17] Chalabianloo N, Can YS, Umair M, et al. Application level performance evaluation of wearable devices for stress classification with explainable AI. *Pervasive Mob Comput*. 2022;87:101703.
- [18] Kim H, Kim M, Park K, et al. Machine learning-based classification analysis of knowledge worker mental stress. *Front Public Health*. 2023;11:1302794.
- [19] Dahal K, Bogue-Jimenez B, Doblas A. Global Stress Detection Framework Combining a Reduced Set of HRV Features and Random Forest Model. *Sensors*. 2023;23:5220.
- [20] Giannakakis G, Grigoriadis D, Giannakaki K, et al. Review on Psychological Stress Detection Using Biosignals. *IEEE Trans Affect Comput*. 2022;13:440–460.

- [21] Matuz A, Van Der Linden D, Darnai G, et al. Generalisable machine learning models trained on heart rate variability data to predict mental fatigue. *Sci Rep*. 2022;12:20023.
- [22] Mueller ST, Piper BJ. The Psychology Experiment Building Language (PEBL) and PEBL Test Battery. *J Neurosci Methods*. 2014; 222:250–259.
- [23] Forte G, Casagrande M. The intricate brain–heart connection: The relationship between heart rate variability and cognitive functioning. *Neuroscience*. 2025;565:369–376.
- [24] Bechara A, Damasio H, Tranel D, et al. The Iowa Gambling Task and the somatic marker hypothesis: some questions and answers. *Trends Cogn Sci*. 2005;9:159–162.
- [25] Kubios Oy. Kubios HRV Scientific User's Guide 2026 [Internet]. Kuopio: Kubios Oy; 2026 [cited 2026 Jan 10]. Available from: <https://www.kubios.com/downloads/HRV-Scientific-Users-Guide.pdf>.
- [26] Laborde S, Mosley E, Thayer JF. Heart rate variability and cardiac vagal tone in psychophysiological study - Recommendations for experiment planning, data analysis, and data reporting. *Front Psychol*. 2017;20:8:213
- [27] Makowski D, Pham T, Lau ZJ, et al. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behav Res Methods*. 2021;53:1689–1696.
- [28] He H, Bai Y, Garcia EA, et al. ADASYN: Adaptive synthetic sampling method for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). Hong Kong, China: IEEE; 2008. p. 1322–1328.
- [29] Radovic M, Ghalwash M, Filipovic N, et al. Minimum redundancy maximum relevance feature selection method for temporal gene expression data. *BMC Bioinformatics*. 2017;18:9.
- [30] Boucsein W. *Electrodermal Activity*. 2nd ed. New York: Springer; 2012.
- [31] Seitz J, Maedche A. Biosignal-Based Recognition of Cognitive Load: A Systematic Review of Public Datasets and Classifiers. In: Davis FD, Riedl R, Vom Brocke J, et al. (eds) *Information Systems and Neuroscience*. Cham: Springer International Publishing; 2022. p. 35–52.
- [32] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM; 2016. p. 785–794.
- [33] Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2:56–67.
- [34] Hörmann T, Hesse M, Christ P, et al. Fine-grained prediction of cognitive workload in a modern working environment by utilizing short-term physiological parameters. In: *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies*. SCITEPRESS - Science and Technology Publications; 2016.4:42–51.
- [35] Delliaux S, Delaforge A, Deharo J-C, et al. Mental Workload Alters Heart Rate Variability, Lowering Non-linear Dynamics. *Front Physiol*. 2019;10:1–14.