Intelligent Human Systems Integration (IHSI), Vol. 200, 2026, 891–900
https://doi.org/10.54941/ahfe1007150

Interpretable Multimodal Framework for Assessing Cognitive Load and Stress in Collaborative Robot Environments

Sandi Baressi Šegota[1], Darko Etinger[1], Ivan Lorencin[1], Nikola Tanković[1], Luka Blašković[1], and Nikola Anđelić[2]

[1]Faculty of Informatics, Juraj Dobrila University of Pula, 52100 Pula, Croatia
[2]Faculty of Engineering, University of Rijeka, 51000 Rijeka, Croatia

ABSTRACT

This study presents an explainable machine learning framework for estimating cognitive load and stress from multimodal physiological and affective data collected during human–robot collaboration tasks. The proposed approach integrates electroencephalography (EEG), electrocardiography (ECG), galvanic skin response (GSR), and emotion-related features with contextual task information to model human cognitive states. Data were preprocessed, standardized, and evaluated using a leave-one-participant-out cross-validation scheme to ensure subject-independent generalization. Bayesian optimization was applied to tune the hyperparameters of non–tree-based models, including support vector regression (SVR) for predicting continuous NASA-TLX scores and a multilayer perceptron (MLP) for classifying discrete stress levels. The regression model achieved an R^2 of 0.98 and a mean absolute error of 0.08, while the classification model obtained an accuracy and F1-score of 0.94. Model interpretability was ensured through SHapley Additive exPlanations (SHAP) analysis, which identified EEG coherence and beta-band activity, ECG LF/HF ratios, and emotion-related indicators such as sadness and confusion as dominant contributors to increased cognitive load and stress. These findings highlight the potential of combining physiological and affective modalities with explainable artificial intelligence for reliable cognitive state assessment. The developed methodology provides a foundation for adaptive robotic systems capable of monitoring and responding to human mental states, thus supporting safer and more efficient collaboration in dynamic operational environments.

Keywords: Cognitive load, Collaborative robotics, Stress detection, Explainable AI, Human-robot collaboration

INTRODUCTION

Human–robot collaboration (HRC) has emerged as one of the key paradigms in modern industrial and service environments, enabling the combination of robotic precision with human adaptability and decision-making. While such systems offer significant improvements in productivity and safety, their success ultimately depends on the ability of the robot to recognize and adapt to the cognitive and emotional states of its human counterpart (Ajoudani et al., 2018). Excessive workload or stress can lead to errors, reduced efficiency, and even safety risks, emphasizing the need for intelligent

Received November 13, 2025; Revised November 28, 2025; Accepted December 13, 2025; Available online February 1, 2026

systems capable of assessing and responding to human mental states in real time. Consequently, the accurate estimation of cognitive load and stress has become a crucial component of user-centered adaptive robotics and human factors engineering (Lu et al., 2022). Physiological signals provide an objective window into cognitive and affective processes, complementing subjective evaluations such as the NASA Task Load Index (NASA-TLX). Electroencephalography (EEG), electrocardiography (ECG), galvanic skin response (GSR), and emotion-derived metrics have all been shown to contain reliable markers of mental workload and stress. However, leveraging these multimodal signals for reliable state estimation remains a methodological challenge (Wu, 2024). Each modality captures a different aspect of the psychophysiological response, and their relationships with cognitive or emotional states are nonlinear and subject-dependent. This complexity calls for modeling approaches capable of capturing nonlinear interactions while remaining interpretable and generalizable across individuals. Traditional statistical methods often fall short in this context due to their limited ability to model complex multimodal dependencies. Conversely, modern machine learning methods can effectively learn such mappings but often sacrifice interpretability—a crucial requirement for safety-critical applications such as collaborative robotics. To address this gap, explainable machine learning techniques, particularly those based on Shapley Additive Explanations (SHAP), have gained traction as tools that allow transparent insight into the contribution of each physiological or affective feature to model predictions. When combined with robust validation schemes and multimodal data fusion, these techniques enable the development of predictive models that are both accurate and interpretable. In this study, a comprehensive modeling framework was developed to predict cognitive load and stress levels from multimodal physiological and affective data collected in an HRC setting. The framework integrates advanced preprocessing, Bayesian-optimized non–tree-based regression and classification models, and model-agnostic interpretability analysis. The primary objectives were (i) to evaluate the predictive performance of different non–tree-based approaches on multimodal physiological inputs, and (ii) to identify the physiological and emotional features most relevant to cognitive load and stress estimation. The remainder of the paper is structured as follows. The next section describes the dataset, preprocessing steps, and the proposed methodology for regression, classification, and SHAP-based explainability. The following section presents and discusses the experimental results, including model performance and feature importance interpretation. The final section summarizes the main findings and contributions of this work, and outlines perspectives for future development of adaptive and explainable human–robot collaboration

METHODOLOGY

The methodology applied in this work integrates multimodal physiological data processing, regression and classification modeling through non–tree-based machine learning techniques, and interpretability assessment using

explainable AI analysis. The complete workflow begins with the preprocessing and normalization of multimodal sensor data, continues with model training and Bayesian hyperparameter optimization under leave-one-participant-out cross-validation, and concludes with model-agnostic interpretability analysis using SHAP. This combination enables both robust predictive performance and transparent insight into the physiological and emotional correlates of cognitive load and stress. The following sections describe the dataset, applied modeling techniques, and explainability procedures in detail.

Description of the Used Data

The SenseCobotFusion dataset was employed as the experimental foundation of this study. It represents a multimodal collection of physiological and emotional parameters obtained during human–robot collaboration experiments, in which participants performed a predefined series of cognitive and manual tasks. The data include recordings from several sensors capturing electroencephalographic (EEG), electrocardiographic (ECG), and galvanic skin response (GSR) signals, complemented by emotion-derived metrics computed from facial expressions and behavioral indicators (Borghi et al., 2024; Borghi et al., 2025). Each sample additionally contains contextual information describing the performed task, as well as the corresponding NASA-TLX subjective workload score reported by the participant. EEG features comprise absolute and relative power spectral densities across canonical frequency bands (theta, alpha, and beta), as well as lateral asymmetry indices and functional coherence values between electrode pairs. ECG features include time-domain measures such as SDNN, RMSSD, and pNN50, frequency-domain measures including very-low-frequency (VLF), low-frequency (LF), and high-frequency (HF) components, and the LF/HF ratio, alongside non-linear metrics such as entropy-based complexity indices. GSR features include amplitude, response rise and recovery times, and spectral band powers reflecting sympathetic activation. Emotion-related features describe the intensity of several affective states such as engagement, joy, fear, anger, and attention, derived from real-time affective inference algorithms. Finally, the categorical variable Task defines the operational context for each observation, while the variable Label corresponds to the NASA-TLX rating representing subjective cognitive load. For predictive modeling, the continuous Label variable was used as the target output in the regression task, while a discretized version—obtained by dividing the continuous scores into three equally populated intervals representing low, medium, and high stress—was used as the classification target. To ensure subject-independent model validation and to avoid leakage between training and testing data, a leave-one-participant-out (LOPO) cross-validation scheme was adopted. Data preprocessing involved the elimination of missing records or, where applicable, median-value imputation, followed by z-score normalization of all numeric variables. Categorical variables were encoded using one-hot encoding, guaranteeing a consistent numerical representation for all modeling techniques.

Regression and Classification Techniques

Two independent learning pipelines were implemented. The first focused on the regression of continuous NASA-TLX values, while the second targeted classification into discrete stress levels. Both pipelines employed non–tree-based algorithms, emphasizing differentiable, probabilistic, or kernel-based formulations that allow smooth parameter spaces and enable the use of gradient-based interpretability methods. The selection of optimal hyperparameters was conducted through Bayesian optimization using the Optuna framework. The optimization process employs a probabilistic search mechanism based on the Tree-structured Parzen Estimator (TPE), which models the likelihood of candidate hyperparameter values leading to improved performance and updates this distribution iteratively as new trials are evaluated. This approach provides an efficient exploration–exploitation balance compared to exhaustive or random search, yielding competitive configurations with a limited number of trials. Each sampled configuration was trained and validated under the LOPO scheme, and its performance was scored using the determination coefficient (R^2) for regression and macro-averaged F1-score for classification. For regression, additional metrics including the mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE) were computed to characterize prediction accuracy, while the classification models were additionally assessed using accuracy, macro-averaged precision, and recall. The regression pipeline explored four principal model types. ElasticNet regression served as the linear baseline, combining L1 and L2 regularization to promote both sparsity and coefficient stability, where the model minimizes the penalized least squares objective:

$$\min_{\beta}\left[\left\|y - X\beta\right\|^2 + \lambda_1\left\|\beta\right\|_1 + \lambda_2\left\|\beta\right\|_2^2\right] \tag{1}$$

Support Vector Regression (SVR) employed an ε-insensitive loss function and radial basis function kernel, enabling nonlinear mappings between physiological features and workload scores. Gaussian Process Regression (GPR) modeled the target function as a stochastic process with a covariance kernel $k(x_i, x_j)$ defined as the sum of a squared-exponential (RBF) and white-noise component, thus providing both mean predictions and predictive uncertainty. Finally, the Multilayer Perceptron (MLP) regressor implemented a fully connected neural architecture with one or two hidden layers containing between 64 and 256 neurons, rectified linear unit activation, and adaptive learning rate control. Regularization was achieved via weight decay (L2 penalty) and early stopping to mitigate overfitting. Bayesian optimization adjusted network size, learning rate, and regularization strength, ensuring convergence toward stable generalization performance. The classification pipeline followed an analogous structure, addressing the three-class stress-level prediction. The model space included Logistic Regression with ElasticNet penalty (Chemlal et al., 2024), Support Vector Classification (SVC) with RBF kernel (Palanivel et al., 2024), Gaussian Process Classification (GPC) (Yang et al., 2024), and MLP classification networks with the same architecture range as in regression but with softmax output layers (Baressi Šegota et al., 2024).

Each model was optimized to maximize the macro-averaged F1-score under LOPO validation. The use of non–tree-based classifiers allowed a continuous decision boundary and differentiable probability surfaces, facilitating subsequent interpretability analysis. When a configuration failed to converge or produced nonfinite metrics, the trial was pruned and excluded from the optimization process, ensuring robustness of the search.

Explainable AI and SHAP Analysis

The final stage of the methodology addressed interpretability through SHAP analysis. The objective of this stage was to quantify the contribution of each input feature to the model's predictions, thereby identifying which physiological and emotional parameters most strongly influenced cognitive load and stress estimation (Ponce-Bobadilla et al., 2025). SHAP values are grounded in cooperative game theory, where each feature is treated as a player contributing to the overall prediction, and its Shapley value represents the average marginal contribution of that feature across all possible subsets of features. Formally, for a model f with input features x_1, x_2, …, x_n, the contribution of features is defined as (Wang et al., 2025):

$$\phi_i = \sum_{S \subseteq (1 \ldots n) \setminus \{i\}} \frac{|S|!\left(n - |S| - 1\right)!}{n!} \left[f\left(s \cup \{i\}\right) - f\left(s\right) \right] \tag{2}$$

Since exact computation of Shapley values is combinatorially expensive, the KernelExplainer implementation was applied. The KernelExplainer approximates these values using weighted linear regression over randomly sampled feature coalitions, where the sampling weights are derived from the Shapley kernel. Specifically, the explainer approximates the local behavior of the model around each instance by repeatedly perturbing feature subsets and evaluating the model's output, effectively constructing a local linear surrogate that reproduces the model's response surface. This procedure provides consistent, model-agnostic attributions for both linear and nonlinear estimators. In the regression case, SHAP analysis was performed directly on the continuous NASA-TLX predictions, yielding global and local explanations of how each feature contributed to the final cognitive load estimate. For the classification models, the SHAP framework was applied to the model's predicted probability of the "High Stress" class, thereby quantifying each feature's effect on increasing or decreasing the likelihood of elevated stress. To maintain computational tractability, SHAP values were computed on a subset of fifty representative samples from the transformed feature space. The resulting SHAP distributions were visualized as summary plots showing the magnitude and direction of each feature's impact across all analyzed samples. Positive SHAP values correspond to an increase in predicted load or stress, whereas negative values indicate a mitigating influence. The interpretation of these attributions enabled the identification of physiologically meaningful patterns across modalities. Features such as increased EEG alpha asymmetry, higher ECG LF/HF ratios, and elevated GSR amplitudes typically exerted positive

SHAP effects, corresponding to heightened stress or cognitive engagement, while indicators associated with relaxation or autonomic recovery, such as stronger HF components or shorter GSR recovery times, showed negative contributions. The integration of SHAP analysis thus provided both global interpretability—clarifying which modalities dominated the predictive process—and local interpretability—illustrating individual participant responses within specific task contexts. The described methodology establishes a complete, interpretable framework for modeling cognitive load and stress from multimodal physiological and affective data. By combining rigorous preprocessing, Bayesian-optimized non–tree-based learning, and model-agnostic SHAP explainability, the approach achieves both high predictive validity and insight into the underlying physiological mechanisms of workload perception. The results obtained from the implemented framework are presented and discussed in the following section.

RESULTS

The obtained results for the best model are shown in Figure 1, and demonstrate strong predictive capabilities in both regression and classification tasks. The Support Vector Regression model achieved an average R^2 of 0.98, indicating that the predicted NASA-TLX scores closely follow the observed values with minimal residual error. The mean absolute error and root mean square error of 0.08 confirm the precision of the model across participants, while the mean absolute percentage error of 4.83% indicates consistent stability even across varying task conditions. Such performance suggests that the selected kernel configuration and optimization strategy were highly effective in capturing nonlinear dependencies between physiological signals and perceived workload. The classification model, based on a multilayer perceptron architecture, achieved high accuracy (0.94) and a macro-averaged F1-score of 0.94, supported by high precision (0.97) and recall (0.91). These metrics imply that the model maintains a strong balance between identifying high-stress conditions and avoiding false alarms, which is critical in stress recognition applications. Overall, both models demonstrate robust generalization under leave-one-participant-out validation, confirming that the multimodal features provide a reliable basis for estimating cognitive load and stress responses across individuals.

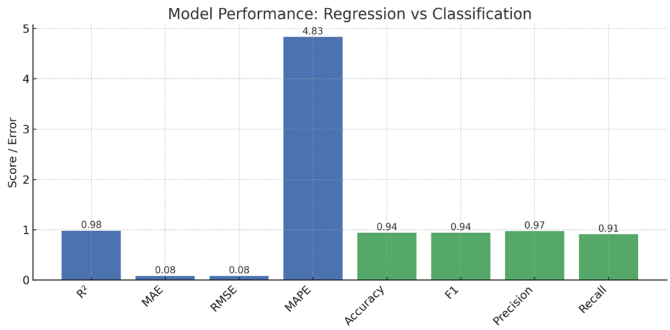

Figure 1: Results of classification (blue) and regression (green).

The SHAP summary plot for regression, in Figure 2, illustrates the relative importance and direction of influence of EEG-derived coherence and spectral features in predicting the continuous NASA-TLX cognitive load scores. The most influential attributes correspond to inter-hemispheric and intra-hemispheric coherence measures, such as P8–T7 and C3–T7 coherence, together with localized alpha and beta band activities. High feature values (red points) of P8–T7 coherence and alpha power tend to increase the model's predicted cognitive load, whereas low values (blue points) generally contribute negatively. This relationship suggests that stronger neural synchronization between posterior and temporal regions, coupled with elevated alpha-band activity, corresponds to higher subjective workload. The inclusion of fractal dimension (Higuchi FD) further implies sensitivity to EEG signal complexity, where increased irregularity also coincides with elevated load conditions. Overall, the regression SHAP profile reveals that distributed cortical connectivity, particularly within temporal–parietal regions, plays a dominant role in continuous workload estimation.

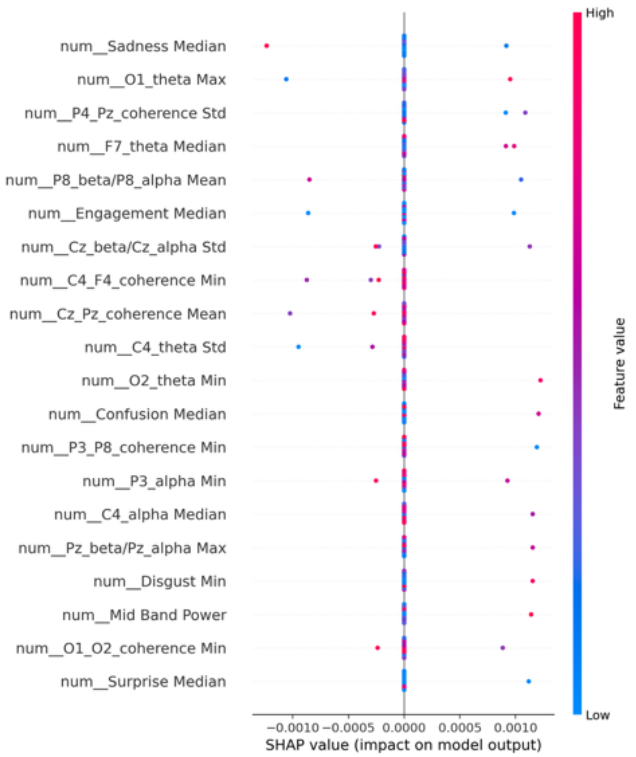

Figure 2: Results of SHAP analysis for the regression model.

In the case of classification, for which the SHAP summary is given in Figure 3, highlights the features most relevant for predicting discrete stress levels. In this case, emotional parameters such as Sadness, Engagement, Confusion, and Surprise appear among the leading contributors, together with frontal and occipital theta-band features and inter-channel coherence ratios. Elevated median values of sadness and confusion, as well as increased

frontal theta activity, positively influence the probability of a high-stress classification, aligning with known associations between frontal theta and cognitive strain. Conversely, higher engagement or increased posterior coherence tends to shift predictions toward lower stress categories. The comparatively smaller magnitude of SHAP values relative to the regression case indicates that the stress classifier relies on a more distributed set of subtle features rather than a few dominant ones, reflecting the inherently more discrete and context-dependent nature of stress-level discrimination.

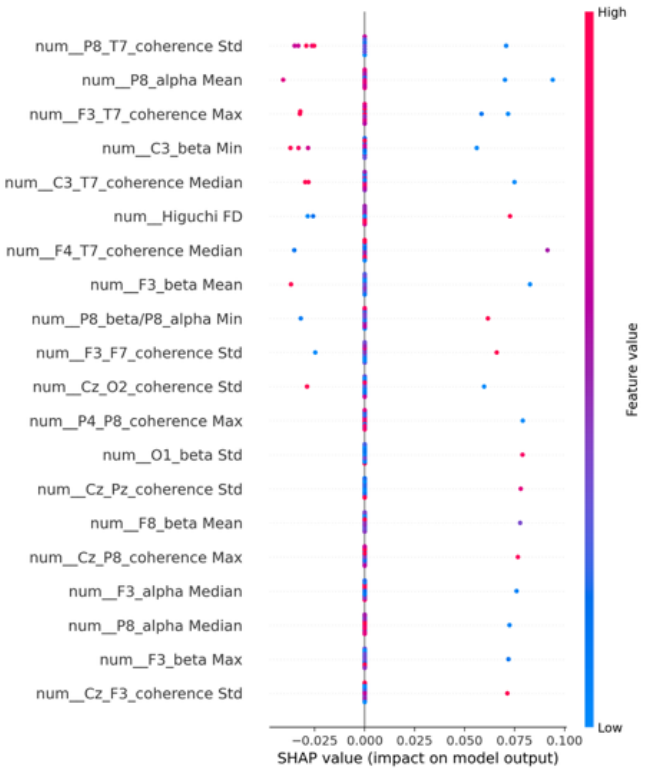

Figure 3: Results of SHAP analysis for the classification model.

CONCLUSION

This study presented an integrated approach for modeling cognitive load and stress using multimodal physiological and affective data collected during human–robot collaboration tasks. The methodology combined rigorous data preprocessing, non–tree-based machine learning models, and model-agnostic explainability techniques. Data from EEG, ECG, GSR, and emotion inference modalities were normalized and aligned per participant and task, allowing the prediction of both continuous NASA-TLX workload scores and discrete stress categories. Bayesian optimization under leave-one-participant-out cross-validation was used to tune hyperparameters and ensure subject-independent generalization. The final models—support vector regression

for continuous cognitive load and a multilayer perceptron for stress-level classification—demonstrated excellent predictive performance, with R^2 = 0.98 and mean absolute error of 0.08 in regression, and an accuracy and F1-score of 0.94 in classification.

Explainability analysis using SHAP revealed that EEG coherence and spectral features, particularly within temporal and parietal regions, exerted the strongest influence on cognitive load estimation, while emotional indicators such as sadness, engagement, and confusion significantly shaped stress classification outcomes. These findings confirm that both neural connectivity patterns and affective cues contribute meaningfully to workload and stress assessment. Overall, the results validate the feasibility of multimodal fusion combined with interpretable machine learning for real-time cognitive state monitoring, providing a foundation for adaptive and user-aware robotic systems capable of responding intelligently to human cognitive and emotional conditions.

ACKNOWLEDGMENT

The authors would like to acknowledge the usage of AI/LLM tools such as ChatGPT and Grammarly for help in text generation and correction.

This research was funded in part by the following projects: SPIN projects IP.1.1.03.0120, IP.1.1.03.0028 and IP.1.1.03.0039; UNIRI projects UNIRI-IZ-25-220 and UNIRI-IZ-25-6; the EU NextGeneration under the Juraj Dobrila University of Pula institutional research project number IIP_010144 and IIP_010136; and EC Digital Europe Programme EDIH Adria 101083838.

REFERENCES

Ajoudani A, Zanchettin AM, Ivaldi S, Albu-Schäffer A, Kosuge K, Khatib O. Progress and prospects of the human–robot collaboration. Autonomous robots. 2018 Jun;42(5):957–75.

Baressi Šegota S, Ključević M, Ogrizović D, Car Z. Modeling of Actuation Force, Pressure and Contraction of Fluidic Muscles Based on Machine Learning. Technologies. 2024 Sep 12;12(9):161.

Borghi S, Nuzzaci A, Peruzzini M, Villani V, Bedogni L. SenseCobotFusion Dataset: Unlocking New Avenues for Stress Detection in Industry 5.0. In 2025 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops) 2025 Mar 1 (pp. 80–85). IEEE Computer Society.

Borghi S, Zucchi F, Prati E, Ruo A, Villani V, Sabattini L, Peruzzini M. Unlocking human-robot dynamics: Introducing SenseCobot, a novel multimodal dataset on Industry 4.0. In Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction 2024 Mar 11 (pp. 880–884).

Chamlal H, Benzmane A, Ouaderhman T. Elastic net-based high dimensional data selection for regression. Expert Systems with Applications. 2024 Jun 15;244:122958.

Lu L, Xie Z, Wang H, Li L, Xu X. Mental stress and safety awareness during human-robot collaboration-Review. Applied ergonomics. 2022 Nov 1;105:103832.

Palanivel R, Basani DK, Gudivaka BR, Fallah MH, Hindumathy N. Support vector machine with tunicate swarm optimization algorithm for emotion recognition in human-robot interaction. In 2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS) 2024 Aug 23 (pp. 1–4). IEEE.

Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtmann S. Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. Clinical and translational science. 2024 Nov;17(11):e70056.

Wang H, Liang Q, Hancock JT, Khoshgoftaar TM. Feature selection strategies: A comparative analysis of SHAP-value and importance-based methods. Journal of Big Data. 2024 Mar 26;11(1):44.

Wu Y, Zhang Y, Zheng B. Workload assessment of operators: Correlation between NASA-TLX and pupillary responses. Applied Sciences. 2024 Dec 20;14(24):11975.

Yang X, Farrokhabadi A, Rauf A, Liu Y, Talemi R, Kundu P, Chronopoulos D. Transfer learning-based Gaussian process classification for lattice structure damage detection. Measurement. 2024 Oct 1;238:115387.