IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 1, February 2026, pp. 469~480
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i1.pp469-480

Classifying mental workload of esports players using machine learning

Aisy Al Fawwaz[1], Osmalina Nur Rahma[1,2], Sayyidul Istighfar Ittaqillah[1,3], Angeline Shane Kurniawan[1], Revita Novianti Putri[1], Richa Varyan[1], Aura Adinda[3], Khusnul Ain[1,2], Rifai Chai[4]
[1]Biomedical Engineering Study Program, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia
[2]Biomedical Engineering Innovation Research Group, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia
[3]Master of Biomedical Engineering, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia
[4]Department of Engineering Technologies, School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Melbourne, Australia

Article Info

Article history:

Received Jun 3, 2024
Revised Aug 6, 2025
Accepted Oct 18, 2025

Keywords:

EDA decomposition
Electrodermal activity
Mental workload
Physiological signal analysis
Temporal confounds

ABSTRACT

Electrodermal activity (EDA) peak counts, derived from both tonic and phasic components, are widely used as physiological proxies for mental workload in cognitively demanding tasks, such as esports. However, their specificity remains uncertain, particularly given potential confounding effect of time-on-task. This study analyzes 92 competitive gameplay sessions from a multimodal esports dataset using three decomposition techniques: convex decomposition (cvxEDA), sparse deconvolution (sparseEDA), and time-varying sympathetic activity (TVSymp). From each method, phasic, and tonic peak counts (TPC), as well as their normalized rates, were extracted. We examined their relationship with self-reported workload through correlation analyses, partial correlations controlling for session duration, and linear mixed-effects models (LMMs). While both peak types exhibited strong positive correlations with gameplay duration (r=0.915 for phasic and r=0.856 for tonic), their association with perceived workload vanished once time was accounted for. Across methods, TVSymp yielded the highest discriminative validity with an area under curve (AUC) of 0.880 in classifying high versus low workload. Machine learning (ML) classifiers trained solely on EDA-based features under a leave-one-subject-out (LOSO) scheme outperformed multimodal models that incorporated heart rate variability (HRV). These results underscore need to disentangle temporal structure from cognitive signals when interpreting EDA and call into question the assumption that EDA peak counts alone reliably encode mental workload across individuals.

Corresponding Author:

Osmalina Nur Rahma
Biomedical Engineering Study Program, Faculty of Science and Technology, Universitas Airlangga
Surabaya, Indonesia
Email: osmalina.n.rahma@fst.unair.ac.id

1. INTRODUCTION

Over the past decade, the exponential growth of competitive electronic sports (esports) has significantly reshaped the landscape of digital interaction and human performance [1]–[3]. Esports is no longer a casual pastime. It now demands sustained cognitive focus, emotional regulation, and rapid decision-making under pressure [4], [5]. These cognitive and affective demands increasingly resemble those encountered in high-stakes domains such as aviation, military operations, and emergency response [6].

In response to this convergence, research in affective computing has shifted its focus toward physiological signals as a means to monitor mental states in real-time [7]–[9]. Electrodermal activity (EDA) has emerged as one of the most studied signals due to its sensitivity to sympathetic arousal and its potential to track subtle changes in cognitive effort during naturalistic tasks [10]. In a previous study involving 96 players across 21 esports matches, we evaluated the predictive potential of features derived from EDA and heart rate variability (HRV) in estimating perceived workload using machine learning (ML) models [11]. Among the extracted features, tonic peak count (TPC) and phasic peak counts (PPC) were obtained through convex decomposition (cvxEDA) and consistently outperformed other indicators, yielding area under curve (AUC) scores above 0.88. These peak-based features provided a compact and interpretable summary of autonomic activity, which made them promising for applications in adaptive systems and real-time decision support.

However, the validity of EDA peaks as indicators of cognitive demand remains unresolved. Longer gameplay sessions naturally accumulate more EDA peaks, raising the possibility that peak-based models may conflate task duration with mental effort. Without properly disentangling temporal confounds, classifiers may overfit to time-on-task effects and misinterpret arousal driven by exposure as workload-induced.

The present study addresses this ambiguity through a comprehensive temporal analysis of EDA signals recorded during 92 competitive gameplay sessions. Using three decomposition techniques: cvxEDA, sparse deconvolution (sparseEDA), and time-varying sympathetic activity (TVSymp), we isolate tonic and phasic components and quantify their peak counts and rates. We then apply statistical controls, multilevel modeling, and ML with leave-one-subject-out (LOSO) validation to determine whether these features reflect genuine cognitive workload or are better explained by session length.

Beyond inferential analysis, we assess the generalizability of peak-derived features using ML classifiers trained under subject-independent validation. We compare unimodal EDA models against multimodal variants that include HRV, evaluating whether adding HRV improves cross-subject prediction or introduces variance that hinders performance. By distinguishing between time-driven and workload-specific physiological signals, this study contributes to the development of affective computing systems that are both interpretable and reliable in dynamic, real-world environments such as esports.

2. METHOD

2.1. Dataset and participant

The study utilized the publicly available esports sensors dataset [12], which contains multimodal physiological recordings from 22 competitive league of legends (LOL) matches involving two teams of five players each, totaling 110 individual session instances. All participants were male, aged 18 to 35 years, and were categorized into two cohorts based on their level of expertise. Professional players reported 5,000 to 10,000 hours of gameplay, while amateurs had between 400 and 1,200 hours.

Each session included synchronized recordings of EDA, heart rate (HR), and post-match self-reported workload, collected via wrist-mounted biosensors in naturalistic, tournament-style environments. Rich metadata accompanies each session, including match duration, player role, calendar day of gameplay, and the order of gameplay within the day. Matches were distributed across four consecutive days, with each day comprising multiple sessions indexed both globally and locally (e.g., match 3 on day 1 and match 6 on day 2), enabling precise temporal alignment.

A quality control protocol was applied to exclude sessions with missing or corrupted EDA, insufficient HR data, or incomplete self-reports. This resulted in a clean analytic cohort suitable for downstream statistical and ML analyses. Post-match workload ratings, collected on a five-point Likert scale, were binarized into high (4-5) and low (1-3) workload labels, which served as the primary target variable. The dataset's temporal structure allows for analysis of fatigue effects, circadian variation, and nested modeling across players and sessions.

2.2. Electrodermal signal decomposition

Raw EDA signals were first denoised using Daubechies-4 wavelet transforms to attenuate motion artifacts and suppress high-frequency noise. We then decomposed the signals into tonic and phasic components using three widely recognized methods: cvxEDA, sparseEDA, and TVSymp. The cvxEDA model, introduced by Greco *et al*. [13] assumes that the observed skin conductance signal *y(t)* can be represented as (1).

$$y(t) = r(t) + (h * p)(t) + \varepsilon(t) \quad (1)$$

Where *r(t)* is the slowly varying tonic component, $p(t) \geq 0$ is a sparse phasic driver, *h(t)* is the canonical impulse response of the sudomotor nerve activity, $*$ denotes convolution, and *ε(t)* is Gaussian noise.

SparseEDA, proposed by Gallego *et al*. [14], also models the EDA signal as the sum of tonic and phasic components but focuses on recovering *p(t)* through non-negative least absolute shrinkage and selection operator (LASSO)-based sparseEDA. The model is formulated as (2).

$$\min_{p \geq 0} \|y - h * p\|^2 + \lambda \|p\|_1 \quad (2)$$

With the tonic component estimated using adaptive polynomial smoothing. This approach is lightweight, efficient, and highly suitable for embedded or wearable applications where interpretability and computational speed are critical.

In contrast, TVSymp, proposed by Quintero *et al*. [15], avoids explicit deconvolution. It instead analyzes the EDA signal in the time-frequency domain using variable-frequency complex demodulation. The sympathetic energy *E(t)* is extracted from the signal's time-varying spectral density *S(t,f)* emphasizing the frequency band associated with sympathetic activation as in (3).

$$E(t) = \int_{f_1}^{f_2} S(t,f)\, df \quad (3)$$

Our selection of these three methods was informed by a benchmark comparison conducted by Veeranki *et al*. [16], which demonstrated substantial variability in decomposition performance across affective computing tasks. Given that esports' settings involve continuous cognitive engagement, externally paced task demands, and minimal gross motor movement, it remains an open question which decomposition approach yields the most behaviorally congruent features under real-time workload. From each decomposition output, we extracted TPC and PPC using a zero-crossing-based local maxima detector implemented via scipy.signal.find_peaks [17]. These features were computed across entire gameplay sessions and within fixed-length temporal segments to support within-session modeling. To determine the most suitable method for workload modeling, we evaluated the statistical correspondence between extracted features and players' self-reported mental workload. Spearman correlation coefficients were computed between TPC, PPC, and their respective normalized rates against binarized workload labels.

2.3. Temporal segmentation and peak rate modeling

To examine whether EDA peak counts exhibit intrinsic temporal accumulation, each session-level EDA signal was segmented into non-overlapping windows of equal duration. Within each segment *s*, we computed the phasic peak rate $R_p(s)$, and tonic peak rate $R_t(s)$, defined as (4).

$$R_p(s) = \frac{P_p(s)}{\Delta t},\ R_t(s) = \frac{P_t(s)}{\Delta t} \quad (4)$$

Where $P_p(s)$ and $P_t(s)$ denote the number of detected phasic and tonic peaks within segment s, and Δt represents the fixed segment duration in minutes.

To assess the presence of time-dependent accumulation, we calculated the Pearson correlation coefficient between the segment index and corresponding peak rates across all segments for each participant. This quantifies the linear association between time elapsed and physiological reactivity. A significantly positive correlation would indicate a systematic increase in EDA peaks over time, independent of subjective workload levels. We additionally computed Spearman rank-order correlations. This dual-metric approach provides a more comprehensive assessment of temporal trends, distinguishing between linear and monotonic accumulation patterns.

2.4. Statistical analysis and temporal disambiguation

To disentangle cognitive workload effects from time-on-task confounds, we employed a multi-tiered inferential framework that incorporated bivariate correlation, partial correlation, linear mixed-effects models (LMMs), and resampling-based validation. We first computed Spearman correlations between total session duration and both phasic and TPC. This analysis assessed whether longer gameplay sessions naturally generate more EDA peaks regardless of cognitive demand, serving as baseline indicator of temporal accumulation.

To isolate the specific association between EDA dynamics and perceived workload, independent of session length, we then computed partial correlations between normalized peak rates (peaks per minute) and binary workload labels, controlling for session duration. This step enabled us to assess whether EDA-derived features retained explanatory power after accounting for temporal exposure. Next, we implemented hierarchical LMMs with phasic and tonic peak rates as dependent variables. Workload was entered as a fixed effect, and random intercepts were specified for player identity and match day to account for interindividual and session-level variability. The base model took the form (5).

$$R_i = \beta_0 + \beta_1 \times W_i + u_{player(i)} + u_{day(i)} + \epsilon_i \quad (5)$$

Where R_i is the peak rate for session i, W_i is the workload condition (0=low, 1=high), and $u_{player(i)}$, $u_{day(i)}$ are random effects. This model was fitted separately for phasic and tonic components.

To investigate potential interaction effects between time and workload, we extended the model to include segment index and its interaction with workload (6).

$$R_{is} = \beta_0 + \beta_1 \times W_i + \beta_2 \times S_s + \beta_3 \times (W_i \times S_s) + u_{player(i)} + \varepsilon_{is} \quad (6)$$

Where S_s is the segment index. A significant interaction term (β_3) would indicate that temporal accumulation patterns differ by workload condition.

Finally, to assess model robustness and generalizability, we conducted a combinatorial resampling analysis. Specifically, we generated 1,000 stratified bootstrap subsets of players and sessions, repeating the full correlation and mixed-model analyses for each. These yielded distributions of parameter estimates (e.g., β_1) from which we derived confidence intervals and stability indices. Such resampling allows us to estimate parameter variability under different sampling configurations and assess the reliability of our inferences across subsets.

2.5. Machine learning classification for mental load with LOSO

To complement our inferential analyses and evaluate the predictive validity of EDA-derived features, we implemented a supervised ML framework using LOSO cross-validation approach [18]. This evaluation protocol was selected to ensure strict generalization to unseen individuals, which is particularly critical in physiological computing, where inter-subject variability can obscure model performance. In each LOSO iteration, data from a single participant was held out as the test set, while models were trained on data from all remaining participants. This process was repeated until each participant had served as the test set exactly once, thereby ensuring comprehensive and unbiased validation across the whole cohort.

Feature vectors were constructed using both physiological and contextual attributes. These included the PPC and TPC and their normalized rates, computed over both the whole session and temporally segmented intervals. Temporal slope coefficients were calculated from peak rate trends over time. We also included HRV features, such as standard deviation of normal-to-normal intervals (SDNN), root mean square of successive differences (RMSSD), coefficient of variation (CV), and Shannon entropy. Session metadata, such as match duration, gameplay order within the day, and player role, were appended to capture situational factors.

All features were standardized using z-score normalization, computed exclusively on the training folds to prevent information leakage into the test set. To model the data, we trained four classifiers, representing both linear and nonlinear paradigms: support vector machines (SVM) with a radial basis function kernel (RBF), a multilayer perceptron (MLP), neural network (NN), decision trees (DT), extreme gradient boosting (XGBoost), and logistic regression (LR) as a baseline. Hyperparameters for each classifier were optimized via nested grid search within each training fold. We computed multiple classification metrics, including accuracy, AUC, F1-score, and logarithmic loss. Accuracy was defined as the proportion of correctly classified instances over the total number of predictions. AUC was estimated by numerically integrating the receiver operating characteristic (ROC) curve using the trapezoidal rule, capturing the model's ability to rank positive instances higher than negative ones irrespective of classification threshold. To penalize overconfident misclassifications, we also reported log loss, defined as (7).

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)] \quad (7)$$

Where $y_i \in \{0,1\}$ is the true label and $\hat{p}_i \in (0,1)$ is the predicted probability of the positive class. To mitigate class imbalance, we applied class weighting inversely proportional to class frequency, and used stratified sampling where applicable.

To assess inter-subject robustness, we tracked participant-level confusion matrices and computed the average ROC curve across LOSO iterations. All modeling procedures were implemented in Python 3.9 using the Scikit-learn and XGBoost libraries. They were executed in a cloud-based environment (Google Colab) to ensure reproducibility.

3. RESULTS

3.1. Dataset descriptives and quality control

The final dataset included 92 player sessions collected from 22 LOL matches, each involving five individual players with complete recordings of EDA and HR. Sessions were retained after applying basic

quality checks to ensure data completeness and signal usability. And the average match duration was 1,357.5 seconds (approximately 22.6 minutes).

3.2. Decomposition method comparison

Figure 1 shows among the methods, TVSymp consistently outperformed both cvxEDA and sparsEDA across all evaluation axes. It yielded the strongest monotonic association between PPC and mental load ratings (Spearman's ρ=0.640), with a comparable linear relationship, as shown by Pearson's r=0.602. As illustrated in Figure 1(a), TVSymp demonstrates superior phasic response detection, capturing distinct peaks during high-arousal periods that are less pronounced in cvxEDA and sparsEDA decompositions. Notably, this suggests that the envelope-extracted response from the Hilbert-transformed bandpass-filtered EDA captures high-frequency EDA dynamics that scale with transient cognitive arousal, an expected hallmark of phasic sympathetic activation during high-intensity gameplay segments. The other methods, while still positively correlated with workload, showed attenuated sensitivity: cvxEDA produced a moderate correlation (ρ=0.562), whereas sparsEDA's signal-load alignment was substantially weaker (ρ=0.228).

In terms of classification between high- and low-workload states, TVSymp again achieved the highest area under the ROC curve (AUC=0.880), reflecting excellent discriminative power. This was supported by a highly significant difference in PPC between workload classes ($p<1e-7$, Mann–Whitney U) and a large magnitude effect size (Cohen's d=-1.27). These findings indicate that TVSymp-derived phasic features are not only statistically robust but also practically meaningful in distinguishing cognitive demand states under gaming conditions. cvxEDA, while slightly less sensitive, still demonstrated strong performance (AUC=0.807, d=1.08), underscoring its continued relevance in semi-controlled environments. In contrast, sparsEDA underperformed relative to both methods (AUC=0.613, d=−0.47), likely due to its reliance on sparsity-driven assumptions, which may be less stable in noisy, real-world datasets.

While phasic components showed consistent trends, tonic components revealed more heterogeneous behavior across methods. Figure 1(b) reveals that tonic decompositions exhibit greater variability across methods, with cvxEDA showing smoother baseline trends compared to the more oscillatory patterns of TVSymp and sparsEDA. Interestingly, cvxEDA exhibited the most potent tonic–load relationship (ρ=0.464, d=-0.74), suggesting that its slow-varying baseline modeling effectively captures residual autonomic shifts, which may be linked to sustained effort or fatigue. TVSymp, despite its phasic strength, produced weaker tonic associations (ρ=0.377, d=-0.47), possibly reflecting its limited ability to isolate low-frequency baseline trends. sparsEDA, again, lagged behind both correlations and effect sizes, which were near the noise floor. Taken together, these findings highlight that TVSymp offers the most robust and discriminative decomposition for real-time mental workload monitoring in esports, particularly when the goal is to track fast, transient shifts in arousal. CVxEDA remains a solid alternative, especially when tonic modulation or interpretability is desired.

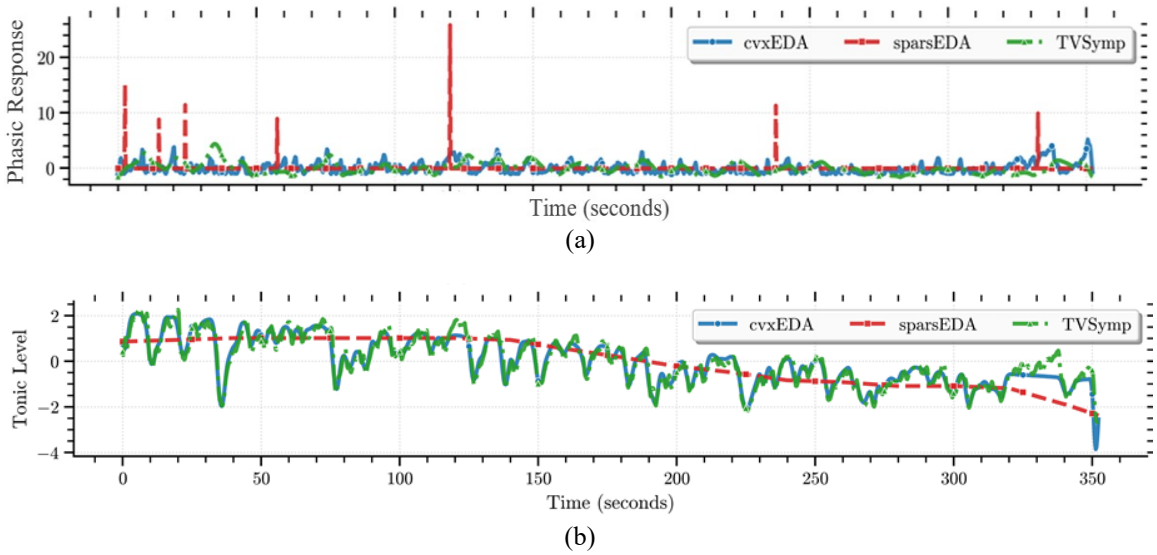

Figure 1. Comparison of EDA signal decomposition methods: (a) phasic and (b) tonic

3.3. Temporal segmentation and trend analysis

Using the TVSymp decomposition method, we analyzed how EDA changes throughout gameplay and its relationship to perceived mental workload. As demonstrated in Figure 2, both phasic and TPC showed strong positive correlations with gameplay duration (r=0.915 and r=0.856, respectively; both ρ<0.001), indicating that sympathetic arousal accumulates steadily during prolonged matches. The linear trend lines clearly illustrate this time-dependent accumulation, with phasic peaks showing a steeper gradient compared to tonic peaks. The increasing phasic activity likely reflects repeated momentary responses to in-game stimuli, while the tonic component may represent slower adaptations such as sustained engagement, stress buildup, or physiological fatigue. The plots in Figure 2 reveal considerable inter-session variability around the regression lines, particularly evident in the tonic component, where data points show greater dispersion at longer durations. This heterogeneity suggests that while the overall temporal trend is robust, individual matches exhibit unique physiological signatures that game dynamics, player strategies, or contextual factors may influence.

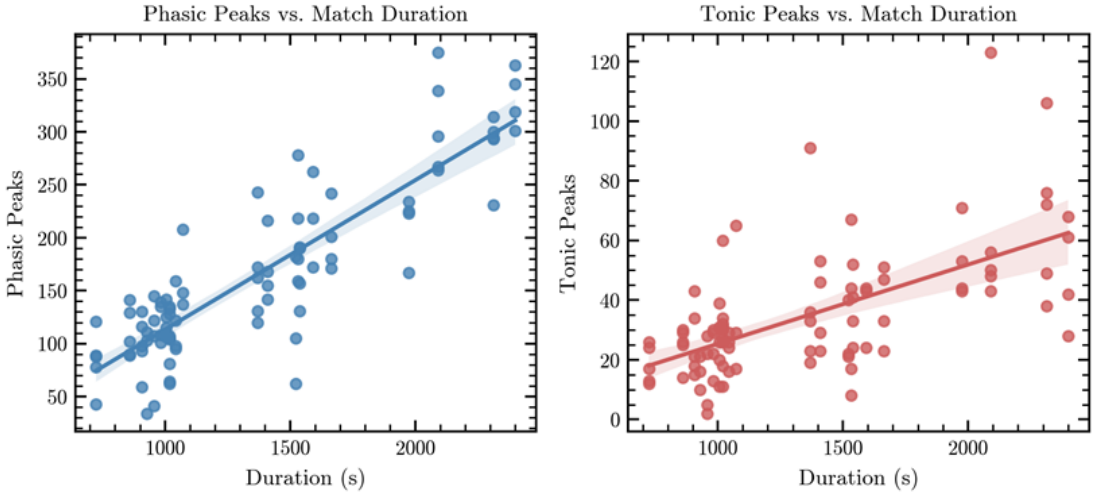

Figure 2. PPC and TPC as a function of match duration. Both show strong positive correlations (phasic: r =0.915, tonic: r =0.856; p <0.001)

However, the relationship between EDA peak rates and subjective workload was weak. Phasic peak rate was only marginally correlated with self-reported mental load (r=0.069, ρ=0.515), and tonic peak rate was effectively uncorrelated (r=-0.023, ρ=0.827). This suggests that physiological arousal does not directly align with coarse match-level workload labels. Several factors may contribute, including the limited granularity of mental load annotations, inter-individual variability in autonomic reactivity, and variations in game pacing or player roles.

3.4. Correlation and partial correlation analysis

Bivariate Spearman correlations revealed strong positive associations between total session duration and EDA peak counts for both phasic (r=0.835, ρ<0.001) and tonic components (r=0.630, ρ<0.001). These findings suggest that as the duration of gameplay increases, the cumulative number of EDA events also tends to rise, possibly reflecting sustained engagement, accumulation of arousal, or fatigue buildup over time. However, when controlling for session duration, the association between normalized peak rates and subjective workload became much weaker. Partial correlations between mental load label and phasic rate (r=-0.092, ρ=0.385) and between workload and tonic rate (r=-0.185, ρ=0.078) were not statistically significant. This suggests that increases in EDA peak density over time may be more attributable to time-on-task than to perceived cognitive demand. The lack of association may also reflect limitations in the granularity of global workload labels, which do not capture intra-match cognitive variability.

3.5. Mixed-effects modeling and bootstrapped validation

To further test the predictive role of workload on EDA dynamics, LMMs were fitted using phasic and tonic rate per minute as dependent variables, with workload as a fixed effect and player as a random

intercept. The model for phasic activity produced a non-significant coefficient (β=0.036, ρ=0.929), indicating no meaningful relationship between workload class and phasic activation rate. Meanwhile, the tonic model yielded a marginally non-significant adverse effect (β=-0.253, ρ=0.093), hinting at a potential inverse association between higher workload and tonic activity. However, this did not meet conventional thresholds for statistical significance.

To assess the robustness of these findings, bootstrap resampling was used to estimate the stability of the phasic model's workload coefficient. Across 1,000 iterations, the mean bootstrapped β_1 was 0.0397, with a 95% confidence interval ranging from -0.8705 to 0.8216. This interval consistently spanned zero, further supporting the interpretation that any workload effect on EDA peak rates is weak and unstable in this dataset.

3.6. Match order, role, and day-level effects on EDA dynamics

Despite expectations that contextual gameplay variables might shape EDA, statistical modeling revealed only modest and inconsistent effects. As shown in Figure 3, phasic peak rates fluctuated across match order and tournament days, with a slight increase observed on day 1 relative to day 0 (β=0.147, p=0.036). However, this pattern was not replicated on day 2, and match order itself did not yield significant effects (p=0.510). Tonic peak rates followed similarly non-significant trends (p=0.601), reinforcing the interpretation that sympathetic activation is not consistently modulated by gameplay sequence or daily progression.

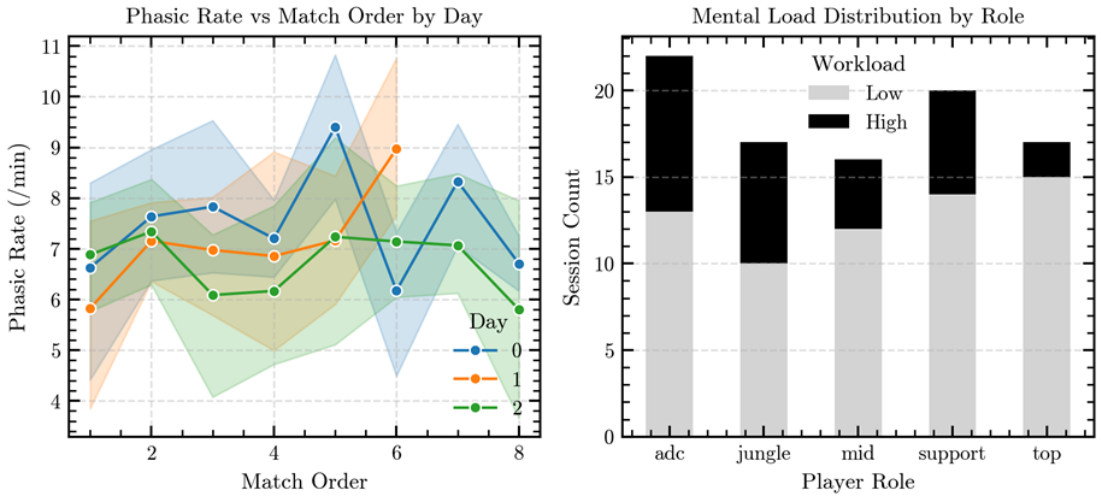

Figure 3. Influence of tournament day, match order, and player role on phasic activation and workload perception

On the right panel of Figure 3, we visualized mental load distribution by player role. While some roles, such as support and attack damage carry (ADC), exhibited higher frequencies of low-load sessions, and jungle and mid roles had slightly more high-load sessions, these differences were not statistically significant. Nonetheless, the role-based trends suggest that task complexity or real-time decision-making demands may subtly influence perceived workload. Taken together, the results suggest that broad contextual features, such as day, match order, or predefined roles, offer limited explanatory power for sympathetic arousal patterns. This underscores the importance of accounting for within-match dynamics or individual differences when modeling EDA responses in competitive settings.

3.6. Machine learning classification

To compare the effects of evaluation strategy and feature selection on model performance, the results from both the previous stratified k-fold setup and the current LOSO validation are summarized in Table 1. Under stratified k-fold, where within-subject samples could appear in both training and testing folds, classification metrics were notably higher across all models. SVM and LR both achieved 81.97% accuracy with an AUC of 0.882, suggesting strong within-subject prediction capabilities. However, these values may overestimate generalizability due to potential data leakage.

Table 1. Classifier performance comparison between stratified k-fold and LOSO using EDA and HRV

Model	Previous study [11] (EDA+HRV, stratified k-fold)	Current study (EDA only, LOSO)	Current study (EDA+HRV, LOSO)
LR	Accuracy: 81.97 AUC: 0.882 F1-score: -	Accuracy: 76.88 AUC: 0.788 F1-score: 0.617	Accuracy: 69.64 AUC: 0.769 F1-score: 0.473
SVM (RBF)	Accuracy: 81.97 AUC: 0.882 F1-score: -	Accuracy: 75.33 AUC: 0.781 F1-score: 0.617	Accuracy: 63.27 AUC: 0.722 F1-score: 0.359
MLP	Accuracy: 82.03 AUC: 0.840 F1-score: -	Accuracy: 70.71 AUC: 0.664 F1-score: 0.437	Accuracy: 65.51 AUC: 0.559 F1-score: 0.283
XGBoost	Accuracy: 79.63 AUC: 0.874 F1-score: -	Accuracy: 69.44 AUC: 0.767 F1-score: 0.482	Accuracy: 67.30% AUC: 0.669 F1-score: 0.462

In contrast, the LOSO framework, which ensures subject-level separation during testing, produced more conservative but arguably more realistic metrics. The best model under these conditions was LR using only EDA peak features, which achieved 76.88% accuracy and an AUC of 0.788. Notably, when HRV features were included alongside EDA, performance declined across all classifiers. For example, the same LR model achieved an accuracy of 69.64% and F1-score of 0.473. Similar degradations were observed in SVM, MLP, and XGBoost, underscoring the possibility that HRV signals, while physiologically meaningful, may introduce subject-specific noise that hinders cross-individual generalization.

4. DISCUSSION

This study examined whether EDA peak counts are more reflective of mental workload or primarily driven by task duration. Across 92 gameplay sessions, we found that both PPC and TPC were significantly correlated with session length. However, once duration was statistically controlled, the association between EDA peak rates and self-reported workload weakened substantially. This pattern suggests that time-on-task is a dominant driver of EDA peak accumulation, and that EDA metrics alone may overestimate cognitive demand if temporal effects are not taken into account.

The results align with prior findings that link sustained sympathetic activation with prolonged task engagement [19], [20]. However, unlike studies that interpret EDA increases as a direct reflection of cognitive effort [21], [22], our approach explicitly separated duration effects through partial correlations and mixed-effects modeling. The analysis showed that phasic and tonic rates did not strongly predict workload labels, indicating that transient EDA fluctuations are more temporally structured than cognitively specific.

We further investigated contextual variables, including gameplay order, tournament day, and player role. The regression models revealed no meaningful impact of match sequence or day progression on EDA peak rates. Although role-based comparisons showed slight variations in workload labels, the differences were not statistically robust. This suggests that inter-session scheduling or gameplay responsibilities have a limited influence on physiological workload markers, and that within-match dynamics may play a more significant role.

To assess predictive utility, we applied four ML classifiers with LOSO cross-validation. When trained using only PPC and TPC, LR achieved the best results, with 76.9% accuracy and an AUC of 0.788. SVM showed similar performance, while MLP and XGBoost models underperformed. Adding HRV features did not improve classification performance. Instead, it led to consistent declines in F1-scores and increased uncertainty across all models.

This finding contrasts with our previous study [11], [23], which used stratified k-fold validation and included all EDA and HRV features. In that setup, SVM achieved an accuracy of 81.97% and an AUC of 0.882. The discrepancy likely stems from the validation strategy. Stratified k-fold allows data from the same subject to appear in both training and testing folds, potentially inflating model performance [23], [24]. In contrast, the LOSO protocol enforces subject-level independence and better reflects real-world generalization scenarios where calibration data may be unavailable.

The drop in performance after including HRV features suggests that these metrics, although physiologically valid, may encode individual-specific traits that reduce cross-subject generalizability. HRV is influenced by various individual-specific factor such as gender, age, and current physical condition, which can lead to variability in different individuals [25], [26]. EDA-derived features, especially TPC and PPC, appear more consistent and subject-invariant, making them better suited for workload modeling in dynamic and heterogeneous populations.

These findings highlight the importance of both feature selection and validation strategy in physiological computing. Compact EDA-based indicators can provide reliable insights into cognitive state without the complexity or variability introduced by multimodal signals. Future work should explore segment-level workload labeling, incorporate behavioral and contextual data, and evaluate adaptive models that personalize predictions without requiring explicit calibration.

5. CONCLUSION

This study demonstrates that EDA peak counts, particularly phasic responses, are more strongly influenced by task duration than by mental workload. While raw peak accumulation aligns closely with gameplay length, their normalized rates show weak and inconsistent associations with self-reported cognitive demand. Even after controlling for duration, both statistical and ML analyses reveal that EDA peaks have limited predictive value for workload classification across individuals. These findings suggest that EDA peak counts are predominantly time-dependent and should be interpreted cautiously as standalone indicators of mental workload. For robust and generalizable modeling, future systems must account for temporal structure and prioritize validation protocols that reflect real-world variability.

ACKNOWLEDGEMENTS

The authors thank the Faculty of Science and Technology, Universitas Airlangga.

FUNDING INFORMATION

The authors thank the Faculty of Science and Technology, Universitas Airlangga, for funding this research under the Airlangga Research Fund (International Research Network) with grant Number 1669/UN3.LPPM/PT.01.03/2023.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Aisy Al Fawwaz	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓			
Osmalina Nur Rahma	✓			✓	✓					✓		✓		✓
Sayyidul Istighfar Ittaqillah			✓			✓			✓				✓	
Angeline Shane Kurniawan						✓			✓				✓	
Revita Novianti Putri						✓			✓				✓	
Richa Varyan						✓			✓				✓	
Aura Adinda						✓			✓				✓	
Khusnul Ain	✓			✓	✓					✓		✓		✓
Rifai Chai					✓					✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

ETHICAL APPROVAL

This study utilized secondary data from the publicly available dataset. Ethical approval and informed consent were obtained by the original authors during data collection. As this research involved only

secondary analysis of publicly accessible data and did not include any direct interaction with human participants, additional ethical approval was not required.

DATA AVAILABILITY

The eSport sensors dataset is provided in csv. format and openly available in Github at https://github.com/smerdov/eSports_Sensors_Dataset.

REFERENCES

[1] B. Watson *et al.*, "Esports and high performance HCI," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, New York, United States: ACM, May 2021, pp. 1–5, doi: 10.1145/3411763.3441313.

[2] I. V. Hilvoorde, "Editorial: esports and digitalization of sports," *Frontiers in Sports and Active Living*, vol. 4, Sep. 2022, doi: 10.3389/fspor.2022.1040468.

[3] D. Ekdahl, I. V. Hilvoorde, Z. A. Rucińska, and S. Ravn, "Editorial: what is esports performance?," *Frontiers in Sports and Active Living*, vol. 6, Dec. 2024, doi: 10.3389/fspor.2024.1538686.

[4] B. T. Sharpe *et al.*, "Reappraisal and mindset interventions on pressurised esport performance," *Applied Psychology*, vol. 73, no. 4, pp. 2178–2199, Oct. 2024, doi: 10.1111/apps.12544.

[5] O. Leis, B. T. Sharpe, V. Pelikan, J. Fritsch, A. R. Nicholls, and D. Poulus, "Stressors and coping strategies in esports: a systematic review," *International Review of Sport and Exercise Psychology*, pp. 1–31, Aug. 2024, doi: 10.1080/1750984X.2024.2386528.

[6] C. Reale *et al.*, "Decision-making during high-risk events: a systematic literature review," *Journal of Cognitive Engineering and Decision Making*, vol. 17, no. 2, pp. 188–212, Jun. 2023, doi: 10.1177/15553434221147415.

[7] I. T. Pavlidis, T. Chaspari, and D. McDuff, "Editorial: special issue on unobtrusive physiological measurement methods for affective applications," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2564–2566, Oct. 2023, doi: 10.1109/TAFFC.2023.3286769.

[8] P. J. Bota, C. Wang, A. L. N. Fred, and H. P. D. Silva, "A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals," *IEEE Access*, vol. 7, pp. 140990–141020, 2019, doi: 10.1109/ACCESS.2019.2944001.

[9] C. Filippini *et al.*, "Automated affective computing based on bio-signals analysis and deep learning approach," *Sensors*, vol. 22, no. 5, Feb. 2022, doi: 10.3390/s22051789.

[10] G. Geršak, "Electrodermal activity - a beginner's guide," *Elektrotehniski Vestnik*, vol. 87, no. 4, pp. 175–182, 2020.

[11] A. Al Fawwaz, O. N. Rahma, K. Ain, S. I. Ittaqillah, and R. Chai, "Measurement of mental workload using heart rate variability and electrodermal activity," *IEEE Access*, vol. 12, pp. 197589–197601, 2024, doi: 10.1109/ACCESS.2024.3521649.

[12] A. Smerdov, B. Zhou, P. Lukowicz, and A. Somov, "Collection and validation of psychophysiological data from professional and amateur players: a multimodal esports dataset," *arXiv:2011.00958*, Aug. 2021.

[13] A. Greco, G. Valenza, A. Lanata, E. Scilingo, and L. Citi, "cvxEDA: a convex optimization approach to electrodermal activity processing," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 1–1, 2016, doi: 10.1109/TBME.2015.2474131.

[14] F. H.-Gallego, D. Luengo, and A. A.-Rodriguez, "Feature extraction of galvanic skin responses by nonnegative sparse deconvolution," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1385–1394, Sep. 2018, doi: 10.1109/JBHI.2017.2780252.

[15] H. F. P.-Quintero, J. P. Florian, Á. D. O.-Cañón, and K. H. Chon, "Highly sensitive index of sympathetic activity based on time-frequency spectral analysis of electrodermal activity," *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 311, no. 3, pp. R582–R591, Sep. 2016, doi: 10.1152/ajpregu.00180.2016.

[16] Y. R. Veeranki, N. Ganapathy, R. Swaminathan, and H. F. P.-Quintero, "Comparison of electrodermal activity signal decomposition techniques for emotion recognition," *IEEE Access*, vol. 12, pp. 19952–19966, 2024, doi: 10.1109/ACCESS.2024.3361832.

[17] P. Virtanen *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020, doi: 10.1038/s41592-019-0686-2.

[18] S. Kunjan *et al.*, "The necessity of leave one subject out (LOSO) cross validation for EEG disease diagnosis," in *Brain Informatics: 14th International Conference, BI 2021*, Virtual event: Springer, Cham, 2021, pp. 558–567, doi: 10.1007/978-3-030-86993-9_50.

[19] H. F. P.-Quintero, J. P. Florian, A. D. O.-Cañón, and K. H. Chon, "Electrodermal activity is sensitive to cognitive stress under water," *Frontiers in Physiology*, vol. 8, Jan. 2018, doi: 10.3389/fphys.2017.01128.

[20] M. Diarra, J. Theurel, and B. Paty, "Systematic review of neurophysiological assessment techniques and metrics for mental workload evaluation in real-world settings," *Frontiers in Neuroergonomics*, vol. 6, pp. 1-19, Apr. 2025, doi: 10.3389/fnrgo.2025.1584736.

[21] E. Lutin, R. Hashimoto, W. D. Raedt, and C. V. Hoof, "Feature extraction for stress detection in electrodermal activity," in *14th International Joint Conference on Biomedical Engineering Systems and Technologies*, Roma, Italy: SCITEPRESS - Science and Technology Publications, 2021, pp. 177–185, doi: 10.5220/0010244601770185.

[22] A. Boffet, L. M. Arsac, V. Ibanez, F. Sauvet, and V. D.-Arsac, "Detection of cognitive load modulation by EDA and HRV," *Sensors*, vol. 25, no. 8, Apr. 2025, doi: 10.3390/s25082343.

[23] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, and D. Scheinost, "Data leakage inflates prediction performance in connectome-based machine learning models," *Nature Communications*, vol. 15, no. 1, pp. 1-15, Feb. 2024, doi: 10.1038/s41467-024-46150-w.

[24] R. P. Fraga, Z. Kang, and C. M. Axthelm, "Effect of machine learning cross-validation algorithms considering human participants and time-series: application on biometric data obtained from a virtual reality experiment," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 67, no. 1, pp. 2162–2167, Sep. 2023, doi: 10.1177/21695067231192258.

[25] E. Ortega and C. J. K. Wang, "Pre-performance physiological state: heart rate variability as a predictor of shooting performance," *Applied Psychophysiology and Biofeedback*, vol. 43, no. 1, pp. 75–85, Mar. 2018, doi: 10.1007/s10484-017-9386-9.

[26] B. S. Tegegne, T. Man, A. M. V. Roon, H. Riese, and H. Snieder, "Determinants of heart rate variability in the general population: the lifelines cohort study," *Heart Rhythm*, vol. 15, no. 10, pp. 1552–1558, Oct. 2018, doi: 10.1016/j.hrthm.2018.05.006.

BIOGRAPHIES OF AUTHORS

Aisy Al Fawwaz received the B.Eng. degree in Biomedical Engineering from Universitas Airlangga, Indonesia in 2024. His research interests include biomedical signal processing, affective computing, and medical imaging. He can be contacted at email: aisy.al.fawwaz-2020@fst.unair.ac.id.

Osmalina Nur Rahma received the S.T. degree in Biomedical Engineering from Universitas Airlangga, Indonesia, in 2013 and the M.Si. degree in Biomedical Engineering from the University of Indonesia, Indonesia, in 2016. Currently, she is a lecturer at the Biomedical Engineering Study Program, Department of Physics, Faculty of Science of Technology, Universitas Airlangga, Indonesia. Her research interests include biomedical instrumentation, signal processing, and rehabilitation engineering. She can be contacted at email: osmalina.n.rahma@fst.unair.ac.id.

Sayyidul Istighfar Ittaqillah received the S.T. degree in Biomedical Engineering from Universitas Airlangga, Indonesia in 2024 and currently became a master student in Biomedical Engineering from Universitas Airlangga specializing in medical instrumentation. He is involved in research on biomedical subject areas with interests in medical signal and image processing, artificial intelligence, and biomaterial engineering. He can be contacted at email: sayyidul.istighfar.ittaqillah-2020@fst.unair.ac.id.

Angeline Shane Kurniawan received the S.T. degree in Biomedical Engineering from Universitas Airlangga, Indonesia in 2024. Her research interests include signal processing and artificial intelligence. She can be contacted at email: angeline.shane.kurniawan-2020@fst.unair.ac.id.

Revita Novianti Putri received the S.T. degree in Biomedical Engineering from Universitas Airlangga, Indonesia in 2024. Currently, she is a biomedical engineer staff at the Electromedical Manufacturing Company in Indonesia. Her research interests include biomedical instrumentation, artificial intelligence, medical image, and signal processing. She can be contacted at email: revitanoviantip@gmail.com.

Richa Varyan received the S.T. degree in Biomedical Engineering from Universitas Airlangga, Indonesia in 2024. Her research interests include biomedical instrumentation and signal processing. She can be contacted at email: richavaryan1919@gmail.com.

Aura Adinda received the S.T. degree in Biomedical Engineering from Universitas Airlangga, Indonesia in 2023 and currently became a master student in Biomedical Engineering from Universitas Airlangga specializing in medical instrumentation. He is involved in research on biomedical subject areas with interests in medical signal and image processing, artificial intelligence, and biomaterial engineering. He can be contacted at email: aura.adinda-2023@fst.unair.ac.id.

Khusnul Ain received the S.T. degree in Nuclear Engineering from Universitas Gadjah Mada, Indonesia, in 1995, an M.Si. degree in Physical Sciences from Universitas Gadjah Mada, Indonesia, in 2002, and a doctoral degree in Engineering Physics from Bandung Institute of Technology, Indonesia, in 2014. Currently, he is a senior lecturer at the Biomedical Engineering Study Program, Department of Physics, Faculty of Science of Technology, Universitas Airlangga, Indonesia. His research interests include biomeasurement, electrical impedance, computational modeling, and bioinstrumentation. He can be contacted at email: k_ain@fst.unair.ac.id.

Rifai Chai received the B. Eng. degree from Krida Wacana Christian University (UKRIDA) Jakarta, Indonesia in 2000 and the Ph.D. degree in Biomedical Engineering from the University of Technology Sydney (UTS), Sydney, Australia in 2014. From 2000 to 2011 he has worked as electronic-product development engineer with companies in Indonesia and Australia. Currently, he is working as senior lecturer with Faculty of Science, Engineering and Technology, Swinburne University of Technology, Melbourne, Australia. His research interests are brain-computer interfaces, rehabilitation, medical technologies and artificial intelligence. Currently, he is a senior member of IEEE and serves as associate editor for Electronics Letters. He can be contacted at email: rchai@swin.edu.au.