Evaluating Multi-Modal Fusion Machine Learning Models For Mental Workload Prediction in Competitive eSports

Udesh Habaraduwa
Yifei Mao

Abstract

In the fast-paced environment of competitive esports, accurately assessing mental workload is crucial for optimizing player performance. This study explores the efficacy of integrating heart rate (HR) and galvanic skin response (GSR) with electroencephalography (EEG) data to improve the classification accuracy of mental workload in League of Legends (LoL) players. Using physiological and in-game data from both amateur and professional players, our hybrid BCI system leverages convolutional neural networks and ensemble models. Results indicate that a uni-modal classifier provides substantial predictive power (outperforming ensembles) and that task conditions (competition against real versus computer-controlled agents) can alter physiological responses.

1 Introduction

The modern world is increasingly characterized by work environments that place considerable demands on our cognitive resources in dynamic, real-time settings (e.g., piloting air-crafts [1], robot-assisted surgery [2] and driving [3]). Consequently, strategies for evaluating mental workload (MWL) from physiological signals alone (e.g., EEG [4] and fNIRS [5]), which tends to be less intrusive to the on-going performance of the task, has received considerable attention [6]. The promise therein is the possibility of tailoring task conditions (e.g., adjusting task difficulty) or monitoring participant behavior to improve performance on a task [7].

A variety of signals measured from a range of physiological response markers (e.g., heart rate, heart rate variability, skin conductance, etc.) have been utilized as an index into MWL during task performance [6] as an alternative to more subjective measures (e.g., self-reports). Recently, as a result of better performance and lower cost (of both compute and sensors), the use of direct measures of neural activity has become more feasible as well, chief among them Electroencephalography [8]. A recent review [9] shows that EEG and cardiac data (heart rate specifically) are amongst the most common measures used for the measurement of MWL. In terms of predicting mental workload, EEG has been shown to be effective with some studies achieving accuracy scores over 90% in classifying MWL [10–12]. Though a relatively fewer proportion of the reviewed studies used skin conductance, it appears to be reliably sensitive to MWL [9].

Given the success of these independent measures, a reasonable next step is to ask if combining them may improve performance in detecting variations in MWL. Indeed, a number of studies have been conducted which combine multiple signals in an attempt to improve performance in modeling MWL [13]. Combining signals (e.g., EEG with HR) has been shown to result in a number of benefits such as improved signal-to-noise ratio, robustness (e.g., against sensor failure or movement artifacts [14]), and performance [13]. The fusion of signals can happen at different levels , at the sensor level (where multiple sensors capture the same signal) all the way up to the decision level (where models combine predictions based on different features for a final decision)[13]. In terms of performance, fusion models have been able to reach classification accuracies of over 80% [15–17] (though results may be lower in the 60% range [14]. Though multi-modal fusion models reviewed in [13] do not appear to out perform single measure systems in general, there is reason to explore in this direction, especially if one aims to deploy a system in the wild [14]. Notably, many of the reviewed studies appear to have been conducted in lab settings where MWL conditions can be reliably and consistently varied along with sensors that capture the same type of physiological signal (e.g., neural activity with EEG and fNIRS [17]). Many real-world situations may result in drastic changes in MWL as a result of random events and with high variance in the levels of elicited physiological response.

A recent addition to the spectrum of task environments characterized by varying degrees of MWL is video games generally but competitive esports (e.g., League of Legends, DOTA 2, and Counter Strike) in particular. Esports environments feature dynamic, fast paced real-time interactions that can place considerable demands on cognitive resources [18], loading on attention, perception, and information processing [19]. Where as athletes in traditional sports have off seasons, esports players compete all year around, training on average for five hours a day [20]. With the understanding that optimally engaged cognitive faculties are crucial for performance at the professional level [19], methods of monitoring the utilization of cognitive resources during training or competition may be useful in preventing failure modes such as player burnout.

Despite the growing popularity of esports around the world, matched by demand for performance gains as in any professional sport, research into the on-going task demands of competitive esports is sparse. Yet the investigation of hybrid BCIs in esports holds considerable potential due to the nature of the gaming environment. The stationarity of esports players during their task execution allows easy collection of multi-modal physiological data [21], as players can comfortably wear multiple sensors to measure their signals, even in non-

experimental settings. Furthermore, the availability of detailed logs from video games enables the synchronization of behavioral measures with in-game data [22]. This provides the possibility to correlate task complexity and difficulty with the mental states and performance of the players, making esports an ideal setting for mental state monitoring experiments aimed at improving player performance. Such hybrid BCI systems for mental workload assessment can be developed with a plethora of physiological signals [23]. In this study, we focus on galvanic skin response (GSR), which measures the change in electrical conductance on the skin over time, and heart rate (HR), as these signals are often recognized as indicators for mental workload [24–27].

In the present study, we aim to extend the present body of literature in MWL monitoring and prediction to the realm of esports. Here in, we employ multimodal fusion at the decision level on three different modalities (EEG, HR and GSR), first training classifiers on individual signals and using the respective model outputs to train an ensemble classifier to investigate our main research question:

> "What are the performance differences, if any, between multimodal and unimodal classifiers in mental workload classification in a competitive esports setting?"

Specifically, we investigate the following:

1. What level of classification accuracy in MWL can be achieved using individual signals alone ?
2. What combinations of EEG, GSR, and HR return best performance?
3. How sensitive are classifiers to underlying confounding factors (we investigate signals derived when players face real opponents versus bots).

To answer these questions, we employ a collection of models shown to be effective in the litreature ([28, 29] on two different representations of EEG data (2-d spectrogram or extracted statistical features) in combination with 1-d GSR and HR signals.

2 Materials and Methods

2.1 Dataset description

2.1.1 League of Legends

The physiological and in-game data of League of Legends (LoL) players, collected by Smerdov et al. [30], serves as the dataset for this study. LoL is a team based strategy game where two teams of five players compete against one another. Each player controls a unique character (called a *champion*) with a set of abilities and navigate through the game's environment as shown in Figures 1 and 2. As players encounter enemy players, they must employ strategies to defeat their opponents, either individually or in coordination with their teammates. The ultimate objective is to destroy the opposing team's base to secure victory. Successful game play, therefore, relies on many factors, including mechanical skill, team synergy, communication, rapid decision making, situational awareness and adaptability to the dynamic game state [31–33]. Due to this demanding and competitive environment, players can experience varying degrees of fluctuations in their physiological responses [30, 32, 34].

The dataset comprises ten participants divided into two teams: one amateur team and one professional team, each consisting of five players. Both teams played 11 matches against either real opponents (through online match making) or AI-controlled bots. A collection of physiological signals were recorded from the players throughout the games (refer [30] for more specifics of the data collection setup and [35] for in-depth details on LoL gameplay[4]). For this study, we will focus on HR, GSR, and EEG signals to predict MWL, considering their potential associations highlighted in the introduction.

To explore possible confounding factors in classification, the data were seperated into two sets: the "all" dataset comprising all available data from Smerdov et al.'s repository, and the "real" dataset which excludes games against bots. We believe that bot games, due to their simplicity, may not facilitate the same increase in physiological response as compared to player vs player games [36]. This difference is further reflected by players reporting significantly lower mental load [30] and experiencing shorter game durations in bot games, as depicted in Figure 5. Given the potential impact on our model's accuracy in predicting MWL, we will utilize both dataset variants for our analysis.

Figure 1: Bird's eye overview of the LOL map. Bottom left and top right corners are the home bases of the two teams. Each team strives to destroy the base of their opponent.

[4]For a quick overview of a LoL game in progress, one can easily be found by searching on YouTube e.g., `https://youtu.be/yo8PtfKeVNU?si=xolhxeKhlOAQA8EF`

Figure 2: League of Legends gameplay showing a player controlled character, a *champion*(yellow bar) in the in-game environment [37]. On the bottom right is the mini-map showing an overview of the entire map. Bottom center shows the abilities and items of the player controlled unit. Other characters shown here are non-player characters in the game. Note that these are different from the *bots* referred to in the present study. Bots take the role of controlling a champion instead of a human player on the opposite team.

2.2 Physiological signals

2.2.1 Galvanic skin response (GSR)

Electrodermal activity (EDA) may be recorded in several forms [38] and has been shown to be a reliable index into a number of underlying cognitive states such as stress [39] and mental workload [40]. The EDA activity available in the present dataset is galvanic skin response (GSR), recorded as the change in electrical resistance (in Ohms) of the skin modulated by the release of sweat[30]. It is measured by placing two electrodes on the skin and continually measuring the level of resistance. The skin conductance is not under conscious control but is responsive to activity of the sympathetic nervous system (i.e., the "fight or flight" system) [41]. Activation of this system induces a release of sweat onto the skin which in turn is recorded as a reduction in resistance [41]. The signals were captured using *Grove GSR sensor*[1] at a sampling rate of 36 Hz [30]. The sensor data corresponds to a 1-d vector of shape (1 × match length). An example is shown in figure 3.

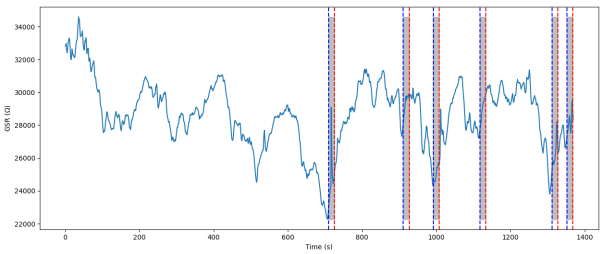


Figure 3: An example time series (GSR) (before normalization) from which windows (shaded) are extracted. Center of the window corresponds to a kill, death, or assist being scored by a given player.

[1] https://wiki.seeedstudio.com/Grove-GSRSensor/

2.2.2 Heart rate (HR)

Cardiac activity has a rich history of being utilized as index into a variety of underlying physiological (e.g., cardiovascular disease [42]) and cognitive functioning [43]. Though heart rate variability (the oscillations of the interval between heart beats) has received most attention in the cognitive functioning domain [43], heart rate (measured as beats per minute) has been shown to be responsive to task demand, number of tasks being carried out, and memory load [6]. Heart rate was captured using the *Polar OHI*[2] armband at a sampling rate of 1 Hz. An example is shown in figure 4.

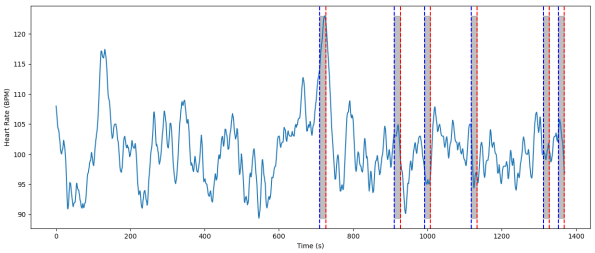


Figure 4: An example time series (HR) (before normalization) from which windows (shaded) are extracted. Center of the window corresponds to a kill, death, or assist being scored by a given player.

2.2.3 Electroencephalography (EEG)

EEG recordings are measurements over time (in micro volts) of the electrical activity that results from the neuronal activity in the cortex of the brain. That is, the neuronal activity generates an electric field, changes in which can be detected by electrodes placed on the scalp [44]. The data used in the present study was generated by recording each player during the entire duration of a match with an Emotive Insight 5-channel EEG headset [30]. The details of the channels recorded are as follows [45] (more details on the Emotiv headset can be found in [46]):

- Pz : Central parietal region
- T8 : Right temporal region
- T7 : Left temporal region
- AF4 : Right anterior frontal region
- AF3 : Left anterior frontal region

From each channel, the activity in five frequency bands is provided (section 2.3.2). An example (for an 8 second window) of the recording can be seen in 6 (note that topographical examples show the average normalized activation).

Since the authors [30] had not made the original raw dataset public, we resolved to use the data provided [3]. The signals used herein were pre-processed by [30]

[2] https://www.polar.com/en/products/accessories/oh1-optical-heart-rate-sensor

to remove outliers, smooth the signals, and resample all signals for a one second time step. Further, only matches in table 8 were included as not all matches had complete recordings of the necessary signals (or signals were missing entirely). The distribution of signal lengths categorized by type of opponents faced are shown in figure 5. In total, the EEG, GSR and HR recordings of 15 matches provided usable data. Summary information for lengths of signals are shown in table 1. Specific lengths can be found in table 8 in the appendix.

Signal	Opponent Type	Mean	SD	Num. signals
EEG	Real	27.92	9.31	21
	Bot	15.57	2.45	13
GSR	Real	31.81	6.20	21
	Bot	16.46	0.98	13
HR	Real	30.33	8.90	21
	Bot	16.47	0.98	13

Table 1: Summary statistics for EEG, GSR, and HR signal length (in minutes) by opponent type.

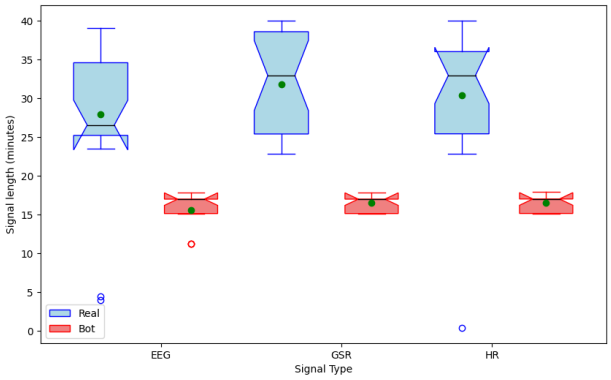


Figure 5: The distribution of signal lengths of the matches included for analysis. Black line indicates median and green dot indicates mean.

2.3 Pre-processing

We considered that each player may display a different baseline level for their physiological responses reflective of individual differences between subjects [47, 48], therefore the mean level for each player was subtracted from their respective signals before further processing. Prior to training the machine learning (ML) models, all GSR and HR data were further scaled, across participants (using StandardScaler [49]), a commonly accepted practice in ML contexts (EEG data was scaled separately, refer section 2.3.2).

[3]We attempted to contact the primary authors for the original dataset but a response was not received.

2.3.1 Generating high-workload and low-workload samples

Given that high-workload or low-workload conditions were not specifically demarcated in the data as would have been done in a traditional experimental setting, it was necessary to derive them. For the present purpose, we define high-workload and low-workload signals as follows:

- **High-workload** : 4 second window around a moment a player was killed, scored a kill , or assisted another teammate in scoring a kill (moments of interest (MOI)) (resulting in an 8 second sample). While in the original dataset also contained a self-report of perceived workload, these were recorded at the end of a match. Notably, the original authors noted a difference in reported scores between matches against real opponents and those against computer bots, with the later being rated as less demanding by the players. Though we considered using these two conditions as the training data for the separate classes, for the present study, we follow the example of [50] and sample around these MOIs in the game where the responses may be most easily be separated from the background activity during game play.

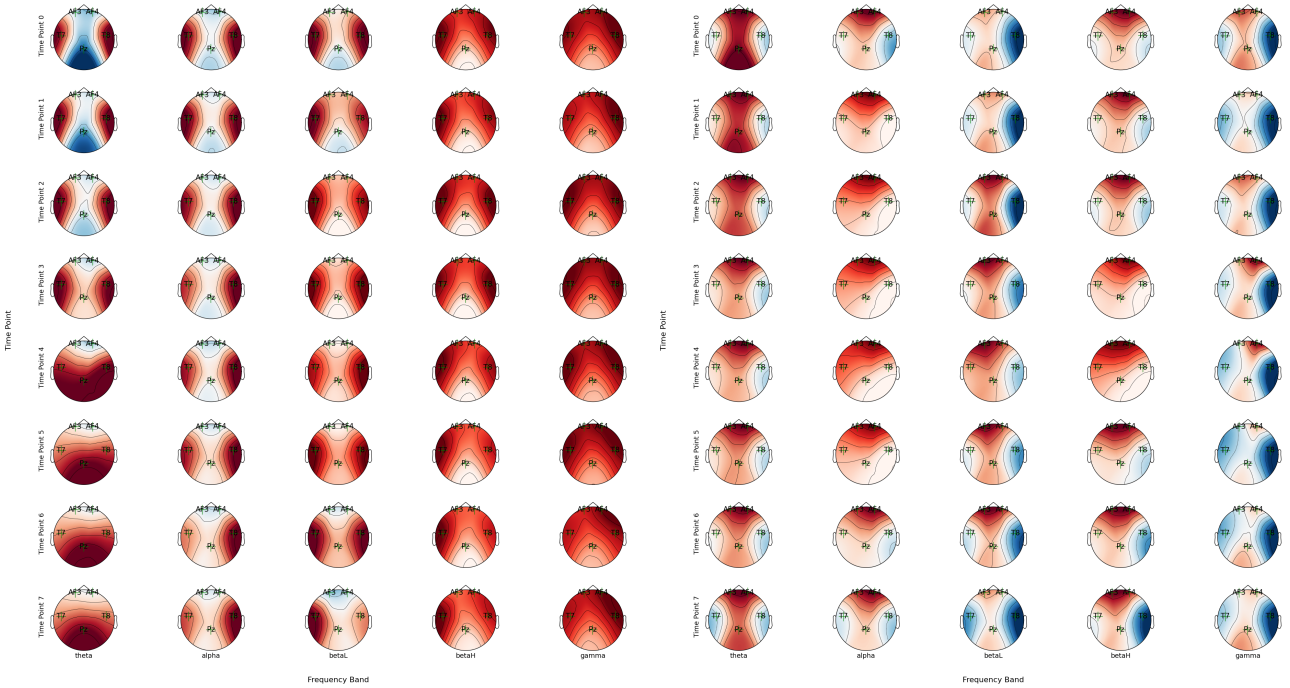
- **Low-workload** : 4 second window (8 second sample) excluding regions that are within the window of a MOI. We sample the same number or fewer than the high-workload windows.

For example, figure 3 shows the windows generated from player three's GSR time series. Topological representations of the positive class (high-workload) and negative class (low-workload) when bot and real data are combined and real only are shown in figure 6 and 7 respectively. Each class contained 637 examples for each signal (in the combined dataset) and 402 examples in the real only dataset. Standardized distribution of values of GSR and HR over an 8 second interval are shown in figure 8 and 9 respectively.

2.3.2 Generating a representations of the EEG data

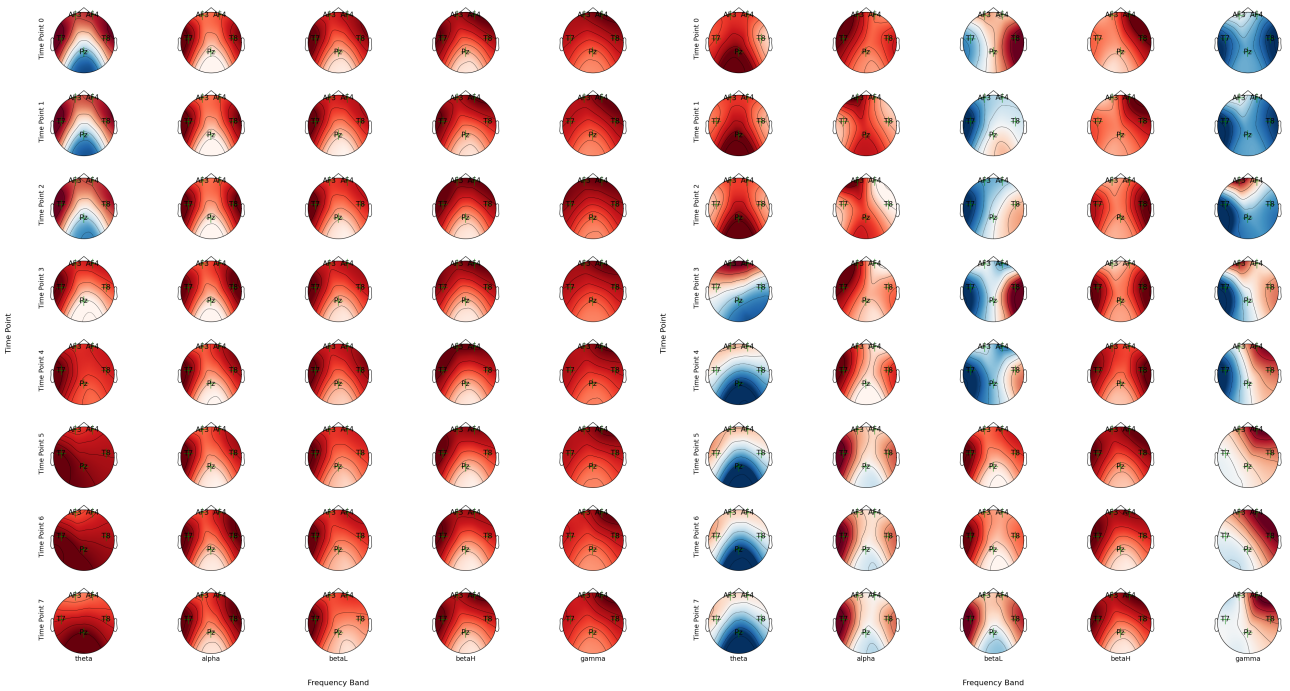
For the present study, we generated two different representations of the EEG data, a spectrogram representation and a statistical representation.

A common approach when subjecting EEG recordings to a ML models is to generate a time-frequency representation of the signal (i.e., a spectrogram) and consequently subjecting the signals to standard computer vision models (e.g., 2-d convolutional networks) [51]. Ideally, these spectrograms are generated from the raw EEG signals such that the horizontal axis represents time t and the vertical axis represents a range of m frequencies. Supposing n channels, for a given period, this would generate n spectrograms, representing the amplitudes of different frequencies in that period (i.e., a matrix of shape ($m \times n \times t$))). The data set



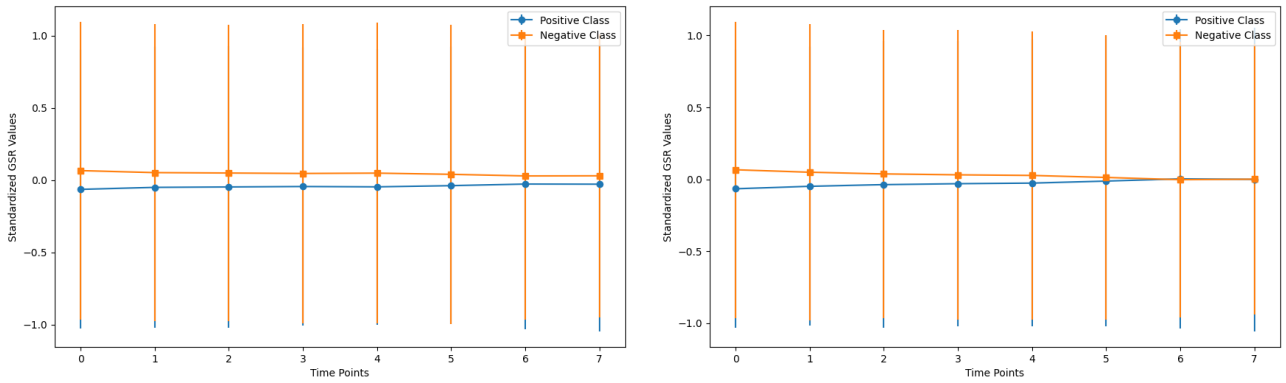
(a) (b)

Figure 6: Topological plots of mean frequency band activity for the positive class (a) and negative class (b) for both bot and real sets combined. Sensor locations on the individual topological map are AF3 (top left), AF4 (top right), T7 (middle left), T8 (middle right), and Pz (bottom).



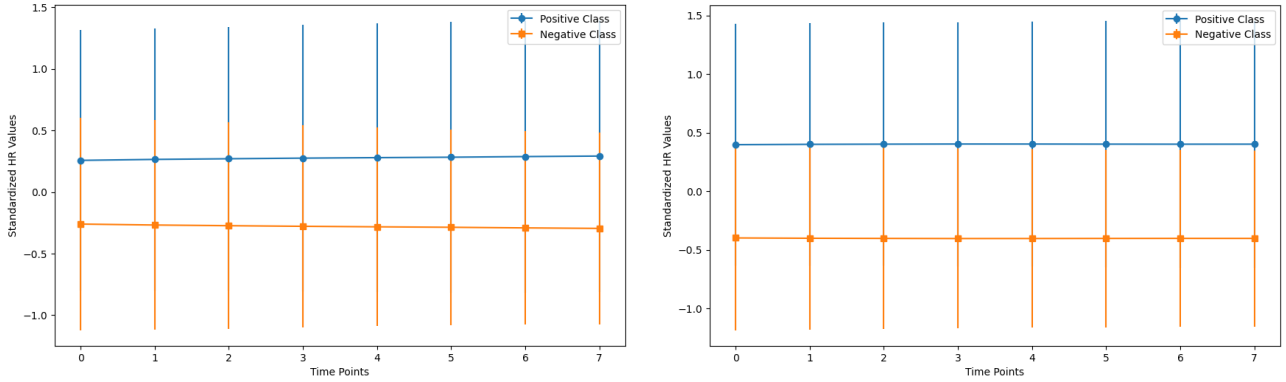
(a) (b)

Figure 7: Topological plots of mean frequency band activity for the positive class (a) and negative class (b) for the real set only. Sensor locations on the individual topological map are AF3 (top left), AF4 (top right), T7 (middle left), T8 (middle right), and Pz (bottom).



(a) (b)

Figure 8: Mean and standard deviation (error bars) of GSR values in the all combined (a) and real only set (b).



(a) (b)

Figure 9: Mean and standard deviation (error bars) of HR values in the all combined (a) and real only set (b).

provides a summarized version of the frequency amplitude. Instead of m frequencies, it provides amplitude activity in frequency bands as follows [52]:

1. Theta (4-8 Hz)
2. Alpha (8-12 Hz)
3. BetaL (Low Beta, 12-15 Hz)
4. BetaH (High Beta, 15-30 Hz)
5. Gamma (30 Hz)

To approximate what may be derived from raw eeg data, we combine the EEG frequency band information and generate a heat map of the values. The matrix is sorted such that the frequency band range increases bottom to top (from theta to gamma). Instead of generating an image for each channel, channels are grouped into 1 image. All values are first scaled by removing the median and scaling according to the quantile range (i.e., robust scaler)[49]. Alternative methods including Z-score and min-max scaler were initially tested but were excluded due to producing excessive amounts of black pixels in the EEG images caused by outliers. An example is shown in figure 10. Note that to improve speed of processing during model training, the EEG spectrogram was decomposed to a single channel gray-scale image and the axes labels and legend are shown here only for reference.

The second representation is a new feature set derived from the statistics of the measured frequency band values [29]. For each frequency band, we calculate the mean, minimum, and maximum value for a given 8 second window. For the given five channels, this transforms each window into a matrix of shape 8 seconds $\times$ (5 channels $\times$ 3 statistics) = (8×15). Distribution of these values is shown in figure 11.

2.4 Machine learning models

2.4.1 Individual models

Our study adapts the neural network architectures proposed by Dar et al. [28], who developed convolutional neural networks (CNNs) to classify emotions based on EEG, GSR, and HR signals. Though the classification target is different, their best performing models achieved accuracies ranging from 90 to 99%, thus motivating our decision to adapt their framework for classifying mental workload. We process GSR and HR data using a 1D CNN with LSTM architecture (Table 2) and EEG data as PNG images through a 2D CNN model (Table 3). All models were implemented with Tensorflow and trained using 100 epochs, a batch size of 32, Adam optimizer, binary cross entropy loss, and

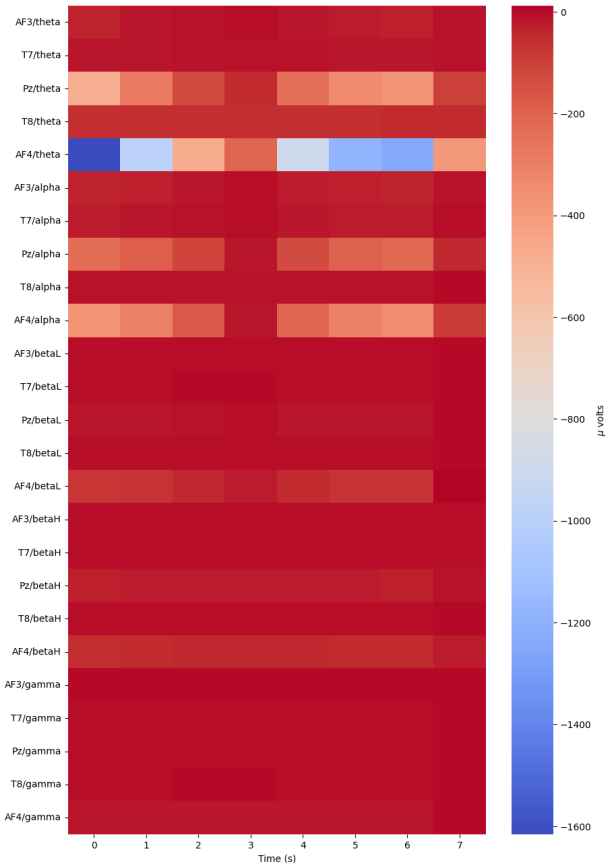


Figure 10: Example of EEG spectrogram. All values are first min-max scaled across the entire dataset. Units are in μ volts.

early stopping (patience of 10).

The second representation of the EEG data as features is trained using random forests and gradient boosting [29] . These models are implemented with the scikit-learn module and optimized by applying cross validation with parameter sweeping.

Table 2: 1D CNN Architecture for HR and GSR Data

Layer Type	Activations	Weights
Input	(8, 1)	-
Conv1D	(8, 16)	3 x 1 x 16
MaxPooling1D	(4, 16)	-
Conv1D	(4, 32)	3 x 16 x 32
MaxPooling1D	(2, 32)	-
Flatten	(64)	-
Reshape	(1, 64)	-
LSTM	(512)	64 x 2048
Dense	(256)	512 x 256
Dropout	(256)	-
Dense	(128)	256 x 128
Dropout	(128)	-
Dense	(4)	128 x 4
Dropout	(4)	-
Sigmoid	(1)	4 x 1

Table 3: 2D CNN Architecture for EEG Data

Layer Type	Activations	Weights
Input	(25, 8, 1)	-
Conv2D	(25, 8, 8)	3 x 3 x 1 x 8
BatchNormalization	(25, 8, 8)	8
MaxPooling2D	(12, 4, 8)	-
Conv2D	(12, 4, 16)	3 x 3 x 8 x 16
BatchNormalization	(12, 4, 16)	16
MaxPooling2D	(6, 2, 16)	-
Conv2D	(6, 2, 32)	3 x 3 x 16 x 32
BatchNormalization	(6, 2, 32)	32
MaxPooling2D	(3, 1, 32)	-
Flatten	(96)	-
Dense	(128)	96 x 128
Dropout	(128)	-
Sigmoid	(1)	128 x 1

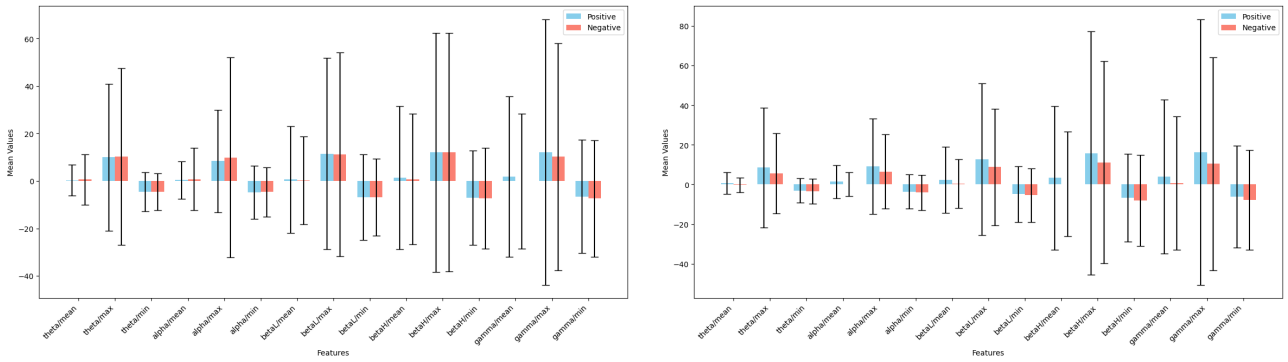
2.4.2 Ensemble models

To address our research question, we evaluate the impact of integrating HR and GSR models with the EEG model on predictive power by creating multiple decision level ensemble models. Each ensemble model contains various combinations of individually trained models that collectively predict a single outcome, as illustrated in Figure 12. The process is guided by the types of input data, which dictate the model configuration. We systematically test combinations in which EEG data was paired with either HR or GSR, or both. Additionally, we test both representations of EEG data independently, ensuring only one EEG representation is used at a time. The individually trained models for EEG, HR, and GSR are saved and provided as input. The ensemble model then integrates these models in the fully connected dense layers where parameters are trained to combine the output of the lower level decision makers. Ultimately, the final dense layer with a sigmoid activation function produces a single probability.

All datasets are divided into 80% train, 10% validation, and 10% test data. The ensemble models requires the data to be aligned in every individual model such that they can be processed collectively for one prediction. This consistency is achieved by applying identical split indices to all datasets, which ensures that the train, test, and validation sets contain the same events for all signals.

To compare the different ensemble models, we perform the training process of 100 epochs for each model 30 times. Subsequently, we perform a one way ANOVA to test for differences in model accuracy. If the ANOVA indicates significant differences, we conduct post-hoc tests using Tukey's HSD to determine which pairwise combinations of models differ from one another.

All models and data can be found at `https://github.com/onedeeper/lolbci`.



(a) (b)

Figure 11: The distribution of means, minimums, and maximums derived for each frequency band for both combined (a) and real-only data (b).

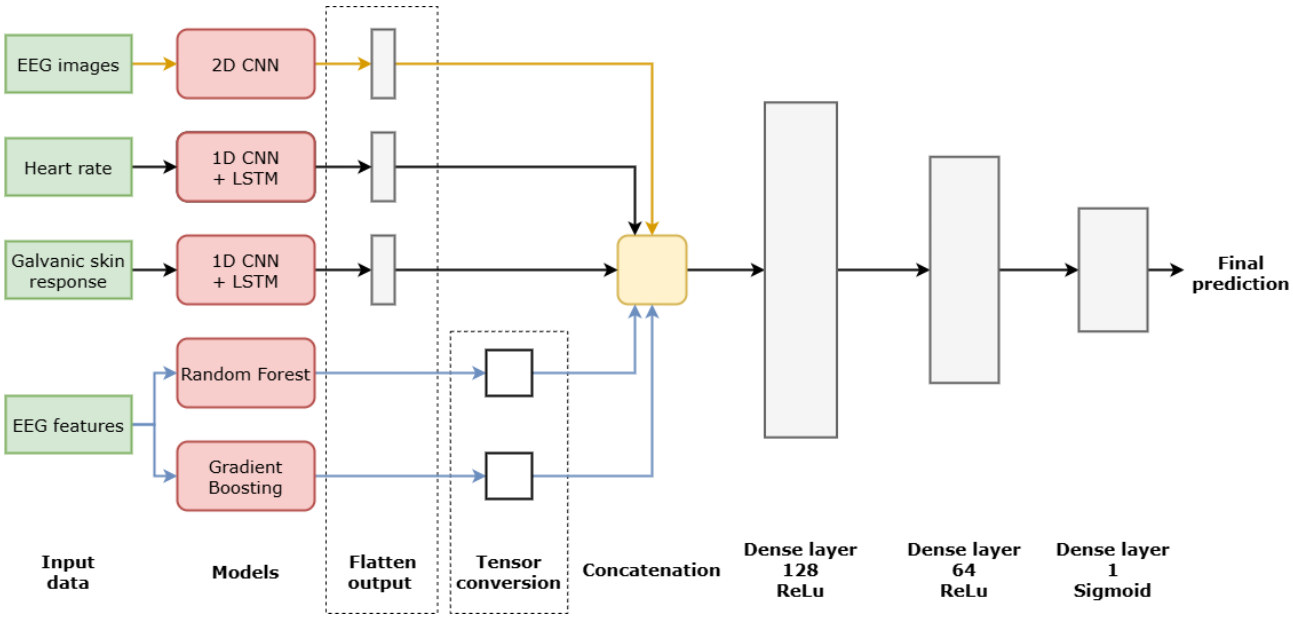


Figure 12: Ensemble models architecture

3 Results

3.1 MWL classification using individual signals

Table 4 presents the accuracy of individually trained models, evaluated on the test set. HR (real) achieves the highest accuracy at (over individual and ensemble models) 75.31%, indicating potential for classifying MWL. This is followed by EEG image (real) with an accuracy of 59.26%, and HR (all) with 57.81%, both of which are just above chance levels.

The other models show much poorer performances. Both GSR models have accuracies around 50%, with GSR (real) at 50.62% and GSR (all) at 53.91%. Models based on EEG features perform even worse, with Random Forest and Gradient Boosting models showing accuracies below chance levels. The highest among these is 48.03% for EEG features (all) using Gradient Boosting while the lowest is 41.25% for EEG features (real) using Random Forest.

Comparing the dataset variants, it can be seen that the real data increases the accuracy of the HR and EEG image models remarkably. Conversely, the GSR and EEG features models show a decrease in accuracy when using the real data. However, this decrease is likely inconsequential, given that these models are not predictive of MWL to begin with.

These results suggest that HR and EEG image can possibly provide meaningful classification accuracies for MWL, especially using the real dataset. On the other hand, GSR and EEG features do not show any predictive power of MWL classification across both dataset variants.

3.2 MWL classification using ensemble models

The ensemble models were trained and evaluated 30 times, with the accuracy recorded for each run. Table 5 presents the mean accuracy over these runs, along with the standard deviation and 95% confidence interval for each model.

To determine if there were significant differences

between the accuracies of the ensemble models, an ANOVA test was conducted. The test returned an F-statistic of 1662 and a p-value of 3.950×10^{-293}, indicating at least one statistically significant difference between the mean accuracy of the models. Subsequently, post hoc Tukey's HSD tests were performed to compare each pair of ensemble models, with the results shown in Tables 6 and 7. Most models differed significantly from other models, with a few of exceptions. Notably, the exceptions arise primarily from comparisons involving the EEG features ensemble models, as all of these report similar accuracies. The only non-significant p-value not involving the EEG features is the comparison between Model 1 and 5. These two ensemble models only differ by the inclusion or exclusion of the individual GSR (all) model.

The highest performing ensemble models are 2 and 6, with accuracies of 70.12% and 71.36% respectively. Both models use HR and EEG image, as well as the real variant of the dataset. The difference is that Model 2 includes GSR in addition to the two signals, which decreases the accuracy slightly.

Similar to the individual features models, all ensemble models using the derived EEG features (3, 4, 9, 10, 11, and 12) also exhibit poor performance with accuracies around 50%. However, the lowest performing model is Ensemble Model 8, with an accuracy of 43.74%. This comes as an unexpected result given that it combines the GSR and EEG image models using the real dataset, each of which individually achieved accuracies of 50% or higher.

Overall, Ensemble Models 1, 2, 5, and 6 demonstrated potential for classifying MWL. The EEG features consistently failed to predict MWL effectively, even when combined with other modalities. GSR combined with EEG also performed poorly (7 and 8). Models incorporating both HR and EEG are the only ones that perform above chance, indicating that this combination of signals is the most useful in classifying MWL out of those considered here.

Comparing the dataset variants show that using the real data significantly improves the classification accuracy by approximately 10%, as evidenced by the comparison between Models 1 and 5 with Models 2 and 6. Thus, the combination of HR and EEG image with the real dataset represents the best ensemble for predicting MWL, achieving an accuracy of 71.36%. However, this accuracy is still lower than that achieved by using HR with real data as an individual model, suggesting that our ensemble process does not enhance the predictive power for MWL.

Table 4: Accuracy of the invididual models

Data	Model	Accuracy
HR (all)	1D CNN + LSTM	57.81%
HR (real)	1D CNN + LSTM	75.31%
GSR (all)	1D CNN + LSTM	53.91%
GSR (real)	1D CNN + LSTM	50.62%
EEG image (all)	2D CNN	52.34%
EEG image (real)	2D CNN	59.26%
EEG features (all)	Random Forest	44.88%
EEG features (real)	Random Forest	41.25%
EEG features (all)	Gradient Boosting	48.03%
EEG features (real)	Gradient Boosting	47.50%

4 Discussion and Conclusion

4.1 Discussion

In the present study, we aimed to extend the hybrid BCI paradigm to the detection of high mental workload states during competitive esports from EEG, HR, and GSR signals. The signals consisted of physiological recordings of ten players engaging in games of League of Legends (LOL). To derive high-workload conditions (positive class), we designated an 8 second window centered around key moments of interest during a match (moments where a player scored a kill, death, or an assist). The negative class was derived as 8 second windows that were not overlapping with the positive class.

First, in comparing the two different representations of the EEG signals (spectrogram versus derived features), we find that the spectrogram representation performs better than the statistical feature models. It may be attributed to the information lost in the aggregated features (minimum, maximum, and mean) but also to the fact that the underlying model for classifying the spectrogram representations (table 3) is vastly more complex (consisting of deep neural networks) than the gradient boosting and random forest models used with the extracted features. Nonetheless, we find this promising and it is conceivable that a higher resolution spectrogram, derived directly from the raw EEG signals would improve performance. Be that as it may, it appears that in the present setup, using EEG alone is not useful in distinguishing between the classes.

Next, we find a difference in performance of the models between the real data vs the all-combined data. Though GSR model performance between the two sets of data are comparable (both performing only at chance levels), both EEG image and HR models perform better when provided with real-only data. Indeed, this suggests that, as noted in the player self-reports [30], that there is a difference in the physiological responses of the players depending on if they play against real opponents or bots. Though we cannot confirm that the players used their personal login details during data collection, if this was the case then the outcomes of the matches against real opponents have immediate consequences (e.g., updating a player's match making ranking [53]), perhaps resulting in a more pronounced response. It also appears this effect between bots versus real players may be present even without these immediate consequences as encounters involving real agents appear to elicit a stronger physiological response regardless [54].

Finally, we find that the ensemble model including

Table 5: Accuracy metrics of ensemble models averaged on 30 runs

Data	Model	Accuracy	Standard Deviation	Confidence Interval
HR, GSR, EEG image (all)	Ensemble Model 1	61.02%	0.84%	[60.71%, 61.32%]
HR, GSR, EEG image (real)	**Ensemble Model 2**	**70.12%**	**2.26%**	**[69.31%, 70.93%]**
HR, GSR, EEG features (all)	Ensemble Model 3	49.32%	0.56%	[49.12%, 49.52%]
HR, GSR, EEG features (real)	Ensemble Model 4	50.58%	0.50%	[50.40%, 50.76%]
HR, EEG image (all)	Ensemble Model 5	60.05%	1.32%	[59.58%, 60.52%]
HR, EEG image (real)	**Ensemble Model 6**	**71.36%**	**1.73%**	**[70.74%, 71.98%]**
GSR, EEG image (all)	Ensemble Model 7	53.85%	1.77%	[53.22%, 54.49%]
GSR, EEG image (real)	Ensemble Model 8	43.74%	1.26%	[43.29%, 44.20%]
HR, EEG features (all)	Ensemble Model 9	49.30%	0.31%	[49.19%, 49.41%]
HR, EEG features (real)	Ensemble Model 10	50.45%	0.89%	[50.14%, 50.77%]
GSR, EEG features (all)	Ensemble Model 11	49.22%	0.00%	[49.22%, 49.22%]
GSR, EEG features (real)	Ensemble Model 12	50.37%	0.74%	[50.11%, 50.64%]

Table 6: Adjusted p-values for pairwise comparisons of ensemble models' accuracy using Tukey's HSD test (n=30). P-values are rounded to three significant figures, with significant values below 0.05 highlighted in bold. (Part 1)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Model 1	*nan*	**<0.001**	**<0.001**	**<0.001**	0.0742	**<0.001**
Model 2	**<0.001**	*nan*	**<0.001**	**<0.001**	**<0.001**	**0.00370**
Model 3	**<0.001**	**<0.001**	*nan*	**0.00290**	**<0.001**	**<0.001**
Model 4	**<0.001**	**<0.001**	**0.00290**	*nan*	**<0.001**	**<0.001**
Model 5	0.0742	**<0.001**	**<0.001**	**<0.001**	*nan*	**<0.001**
Model 6	**<0.001**	**0.00370**	**<0.001**	**<0.001**	**<0.001**	*nan*
Model 7	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Model 8	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Model 9	**<0.001**	**<0.001**	1.000	**0.00210**	**<0.001**	**<0.001**
Model 10	**<0.001**	**<0.001**	**0.0132**	1.000	**<0.001**	**<0.001**
Model 11	**<0.001**	**<0.001**	1.000	**<0.001**	**<0.001**	**<0.001**
Model 12	**<0.001**	**<0.001**	**0.0325**	1.000	**<0.001**	**<0.001**

HR and EEG (images) on the real only dataset out performs all other compared ensembles. Adding GSR to the mix causes a (statistically significant) decrease in the performance. The HR model using the real data outperforms all other models, including the best ensemble model. Indeed, given the poor performance of the base-level GSR models, it is likely that the output provided by this lower level decision maker does not provide a useful signal. The superior performance of the models using HR may be taken to suggest that EEG signals are not useful at all in the present setting and that the improvement over EEG alone is provided entirely by the HR signal alone, though this may be a limitation of the data set (discussed next). Nonetheless, the present results appear to follow the pattern of showing that ensemble models generally do no out perform single signal models [13].

4.2 Limitations and future directions

The findings of the present study suggests EEG signals (individually or as part of an ensemble) are not useful for classifying MWL, despite some evidence to the contrary [10–12]. However, this may be a function of the EEG data that was available. The EEG data set, though rich in many ways, falls short of ideal in the mere fact that they are not the raw collected signals but were subject to a range of pre-processing steps [30]. This drastically limits the resolution of the signals and the objective-specific transformations we could make. We believe that more can be learned by applying the techniques used here by first deriving the necessary information from the raw recorded signals (e.g., a higher resolution spectrogram). In terms of generating new features from the time series data, employing metrics from the the complex dynamical systems literature (e.g., sample entropy [55] and recurrence quantification [56, 57] features) may be useful. It should also be noted that the underlying CNN architectures may not be ideally suited for this task as well (as they were originally intended for emotion classification) and a different set up, perhaps using more recent methods for sequential data (e.g., transformers and attention), may be more effective. In terms of utilizing the existing architecture as is, another possibility is performing an end-to-end training of the ensembles instead of first training base models individually and fusing at the decision level [58].

Considering alternative image-based representations

Table 7: Adjusted p-values for pairwise comparisons of ensemble models' accuracy using Tukey's HSD test (n=30). P-values are rounded to three significant figures, with significant values below 0.05 highlighted in bold. (Part 2)

	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12
Model 1	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Model 2	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Model 3	**<0.001**	**<0.001**	1.000	**0.0132**	1.000	**0.0325**
Model 4	**<0.001**	**<0.001**	**0.00210**	1.000	**<0.001**	1.000
Model 5	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Model 6	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Model 7	*nan*	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Model 8	**<0.001**	*nan*	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Model 9	**<0.001**	**<0.001**	*nan*	**0.00980**	1.000	**0.0247**
Model 10	**<0.001**	**<0.001**	**0.00980**	*nan*	**0.00370**	1.000
Model 11	**<0.001**	**<0.001**	1.000	**0.00370**	*nan*	**0.0102**
Model 12	**<0.001**	**<0.001**	**0.0247**	1.000	**0.0102**	*nan*

of EEG signals may also be fruitful (e.g., topological representations of frequency band activity [59]). Indeed, visually inspecting the two classes (figures 6 and 7) we see noticeable differences in channel activity which appears to support empirical findings [60–62]. Training models specifically on the bands implicated in being reflective of mental workload may also improve discriminability. It may also be useful to represent the EEG images as an n-d array of shape (num. frequency bands × num. channels × seconds). Here we use a 2-d representation, placing all channels on one image which may subtly shift the representations learned by the CNN [63].

Though the performance of even the best performing models appear on the face of it to be relatively poor compared to the referenced literature, we find that these results are nonetheless promising. For one, they suggest that the task conditions (e.g., real versus bots) can impact how the underlying physiology responds, underscoring that caution is necessary when placing these systems in the wild where unforeseen confounds could emerge. Most studies that attempt mental workload classification (e.g., [64], [65], and [1]) are highly controlled experimental settings where the subjects are exposed to precisely timed and graded increases in task demand to vary the dependent variable (mental workload). The present setting is vastly more dynamic with combat situations resulting in a kill, death, or assist occurring seemingly at random. Further, these moments could include just two players or the entire five-person team attacking a single opponent. These are but two of the many dimensions along which workload conditions can vary, all of which may conceivably impact performance. While this may be seen as a limitation of the present work, it is nonetheless illustrative of conditions that a passive BCI (as an ensemble of signals or not) would have to manage in the wild. Thus, the present work, barring the limiting conditions of the data, may actually be more reflective of what we can expect from these systems in more ecologically valid settings.

Finally, all else in the task conditions being equal, there may be a need to account for individual differences within subjects over time and between subjects [66] in these signals when training the models. Indeed, a player's responses may change, for example, as a function of fatigue [67] or professional players may differ in their responses compared to amateurs [55]. All of these may effect how one example is similar to another in the same class. If the statistical properties between examples of the same class differ, it may make learning of the patterns difficult [68] (indeed, we found that models had considerable difficulty during training measured as training accuracy). Thus, testing for stationarity of the signals over time (and correction) and controlling difference between players would be a next step.

In conclusion, in light of the potential benefits of a multi-modal approach [13], we believe that the role of hybrid BCI for monitoring cognitively demanding tasks in the real-world warrants further investigation, especially in domains where neural signals can be collected with minimal interference to the on-going task (like esports). We may also have found a new promising target (HR) as a measure of real-time online changes of MWL. Though we consider only HR and GSR in the present study to augment EEG signals, the space of complementary signals is vast (for example the dataset used herein includes pupil dilation, motion sensor data, etc.) and other signals or combinations there in may result in more performant models. This opens up a wide range of potential sources of information to be integrated to derive a clearer picture of the mind at work.

5 Acknowledgements

5.1 Technology & generative tools statement

GPT-4 and GPT-4o by OpenAI was used for the following tasks:

1. Code for plotting functions in matplotlib
2. Debugging error messages during development
3. Preparing tables for the report
4. Gradient boosting and random forest code
5. Gradient boosting and random forest conversion into tensor code
6. MNE topological plotting code
7. Generating the abstract
8. Miscellaneous help with debugging TensorFlow code

References

[1] Michel De Rivecourt et al. "Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight". In: *Ergonomics* 51.9 (2008), pp. 1295–1319.

[2] Renáta Nagyné Elek and Tamás Haidegger. "Non-technical skill assessment and mental load evaluation in robot-assisted minimally invasive surgery". In: *Sensors* 21.8 (2021), p. 2666.

[3] Henrik Wiberg et al. "Physiological responses related to moderate mental load during car driving in field conditions". In: *Biological psychology* 108 (2015), pp. 115–125.

[4] Anne-Marie Brouwer et al. "Estimating workload using EEG spectral power and ERPs in the n-back task". In: *Journal of neural engineering* 9.4 (2012), p. 045008.

[5] Erin Solovey et al. "Brainput: enhancing interactive systems with streaming fnirs brain input". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 2012, pp. 2193–2202.

[6] Rebecca L Charles and Jim Nixon. "Measuring mental workload using physiological measures: A systematic review". In: *Applied ergonomics* 74 (2019), pp. 221–232.

[7] Christian Mühl, Camille Jeunet, and Fabien Lotte. "EEG-based workload estimation across affective contexts". In: *Frontiers in neuroscience* 8 (2014), p. 83439.

[8] Victoria Peterson et al. "A feasibility study of a complete low-cost consumer-grade brain-computer interface system". In: *Heliyon* 6.3 (2020).

[9] Da Tao et al. "A systematic review of physiological measures of mental workload". In: *International journal of environmental research and public health* 16.15 (2019), p. 2716.

[10] Maarten A Hogervorst, Anne-Marie Brouwer, and Jan BF Van Erp. "Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload". In: *Frontiers in neuroscience* 8 (2014), p. 82981.

[11] Somayeh B Shafiei et al. "Evaluating the mental workload during robot-assisted surgery utilizing network flexibility of human brain". In: *IEEE Access* 8 (2020), pp. 204012–204019.

[12] Matthias Schultze-Kraft et al. "Unsupervised classification of operator workload from brain signals". In: *Journal of neural engineering* 13.3 (2016), p. 036008.

[13] Essam Debie et al. "Multimodal fusion for objective assessment of cognitive workload: A review". In: *IEEE transactions on cybernetics* 51.3 (2019), pp. 1542–1555.

[14] Abhishek Tiwari et al. "Movement artifact-robust mental workload assessment during physical activity using multi-sensor fusion". In: *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. 2020, pp. 3471–3477.

[15] David Rozado and Andreas Dunser. "Combining EEG with pupillometry to improve cognitive workload detection". In: *Computer* 48.10 (2015), pp. 18–25.

[16] Felix Putze et al. "Hybrid fNIRS-EEG based classification of auditory and visual perception processes". In: *Frontiers in neuroscience* 8 (2014), p. 373.

[17] Emily BJ Coffey, Anne-Marie Brouwer, and Jan BF van Erp. "Measuring workload using a combination of electroencephalography and near infrared spectroscopy". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 56. 1. Sage Publications Sage CA: Los Angeles, CA. 2012, pp. 1822–1826.

[18] Alyona Grushko et al. "Perceptual-cognitive demands of esports and team sports: A comparative study". In: *Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics: Proceedings of the 9th International Conference on Cognitive Sciences, Intercognsci-2020, October 10-16, 2020, Moscow, Russia 9*. Springer. 2021, pp. 36–43.

[19] Anna Lisa Martin-Niedecken and Alexandra Schättin. "Let the body'n'brain games begin: toward innovative training approaches in esports athletes". In: *Frontiers in psychology* 11 (2020), p. 138.

[20] Tuomas Kari and Veli-Matti Karhulahti. "Do e-athletes move?: a study on training and physical exercise in elite e-sports". In: *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)* 8.4 (2016), pp. 53–66.

[21] Aaron Koshy et al. "An observation of common physiological parameters during esports activity". In: *International Journal of Esports* 1.1 (2020).

[22] Marçal Mora-Cantallops and Miguel-Ángel Sicilia. "MOBA games: A literature review". In: *Entertainment Computing* 26 (2018), pp. 128–138. ISSN: 1875-9521. DOI: https://doi.org/10.1016/j.entcom.2018.02.005.

[23] Gert Pfurtscheller et al. "The hybrid BCI". In: *Frontiers in neuroscience* 4 (2010), p. 1283.

[24] William Romine et al. "Toward Mental Effort Measurement Using Electrodermal Activity Features". In: *Sensors (Basel, Switzerland)* 22.19 (2022), p. 7363. DOI: 10.3390/s22197363. URL: https://doi.org/10.3390/s22197363.

[25] Roger Lew et al. "Assessing Mental Workload from Skin Conductance and Pupillometry using Wavelets and Genetic Programming". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 54. 2010, pp. 254–258. DOI: 10.1177/154193121005400315.

[26] Sebastian Mach et al. "Assessing mental workload with wearable devices – Reliability and applicability of heart rate and motion measurements". In: *Applied Ergonomics* 105 (2022), p. 103855. ISSN: 0003-6870. DOI: https://doi.org/10.1016/j.apergo.2022.103855. URL: https://www.sciencedirect.com/science/article/pii/S0003687022001788.

[27] Debashis Das Chakladar and Partha Pratim Roy. "Cognitive workload estimation using physiological measures: a review". In: *Cognitive Neurodynamics* (2023). DOI: 10.1007/s11571-023-10051-3.

[28] Muhammad Najam Dar et al. "CNN and LSTM-Based Emotion Charting Using Physiological Signals". In: *Sensors* 20.16 (2020), p. 4551. DOI: 10.3390/s20164551. URL: https://doi.org/10.3390/s20164551.

[29] Nikita Melentev et al. "eSports players professional level and tiredness prediction using EEG and machine learning". In: *2020 IEEE SENSORS*. IEEE. 2020, pp. 1–4.

[30] Anton Smerdov et al. "Collection and validation of psychophysiological data from professional and amateur players: a multimodal esports dataset". In: *arXiv preprint arXiv:2011.00958* (2020).

[31] Lasse Juel Larsen. "The play of champions: Toward a theory of skill in eSport". In: *Sport, ethics and philosophy* 16.1 (2022), pp. 130–152.

[32] Amy X. Zhang and Parth Naidu. *The SIDO Performance Model for League of Legends*. 2024. DOI: 10.48550/arXiv.2403.04873. eprint: 2403.04873.

[33] Júlia Gisbert-Pérez et al. "Key structure and processes in esports teams: A systematic review". In: *Current Psychology* 43 (Mar. 2024). DOI: 10.1007/s12144-024-05858-0.

[34] Anton Smerdov et al. "Detecting video game player burnout with the use of sensor data and machine learning". In: *IEEE Internet of Things Journal* 8.22 (2021), pp. 16680–16691.

[35] Scott Donaldson. "Mechanics and metagame: Exploring binary expertise in League of Legends". In: *Games and Culture* 12.5 (2017), pp. 426–444.

[36] Ken Watanabe et al. "The effects of competitive and interactive play on physiological state in professional esports players". In: *Heliyon* 7.4 (2021), e06844. DOI: 10.1016/j.heliyon.2021.e06844.

[37] Will Burton. *How to Play 1v2 Bot Lane (ADC Guide)*. https://mobalytics.gg/blog/lol-1v2-lane-adc-guide/. Accessed: 2024-07-03. 2023.

[38] DS Bari et al. "Electrodermal responses to discrete stimuli measured by skin conductance, skin potential, and skin susceptance". In: *Skin Research and Technology* 24.1 (2018), pp. 108–116.

[39] Raquel Martinez et al. "A self-paced relaxation response detection system based on galvanic skin response analysis". In: *IEEE Access* 7 (2019), pp. 43730–43741.

[40] Nargess Nourbakhsh et al. "Detecting users' cognitive load by galvanic skin response with affective interference". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7.3 (2017), pp. 1–20.

[41] Mahima Sharma, Sudhanshu Kacker, and Mohit Sharma. "A brief introduction and review on galvanic skin response". In: *Int. J. Med. Res. Prof* 2.6 (2016), pp. 13–17.

[42] Kim Fox et al. "Resting heart rate in cardiovascular disease". In: *Journal of the American College of Cardiology* 50.9 (2007), pp. 823–830.

[43] Giuseppe Forte, Francesca Favieri, and Maria Casagrande. "Heart rate variability and cognitive function: a systematic review". In: *Frontiers in neuroscience* 13 (2019), p. 710.

[44] Fernando Lopes da Silva. "EEG and MEG: relevance to neuroscience". In: *Neuron* 80.5 (2013), pp. 1112–1128.

[45] Oriano Mecarelli. "Electrode placement systems and montages". In: *Clinical Electroencephalography* (2019), pp. 35–52.

[46] Chean Khim Toa, Kok Swee Sim, and Shing Chiang Tan. "Emotiv Insight with Convolutional Neural Network: Visual Attention Test Classification". In: *Advances in Computational Collective Intelligence: 13th International Conference, ICCCI 2021, Kallithea, Rhodes, Greece, September 29–October 1, 2021, Proceedings 13*. Springer. 2021, pp. 348–357.

[47] Federica Cugnata et al. "Modeling physiological responses induced by an emotion recognition task using latent class mixed models". In: *Plos one* 13.11 (2018), e0207123.

[48] Saleh Kalantari et al. "Comparing physiological responses during cognitive tests in virtual environments vs. in identical real-world environments". In: *Scientific Reports* 11.1 (2021), p. 10227.

[49] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[50] Amin Noroozi et al. "An efficient machine learning approach for extracting eSports players' distinguishing features and classifying their skill levels using symbolic transfer entropy and consensus nested cross-validation". In: *International Journal of Data Science and Analytics* (2024). DOI: 10.1007/s41060-024-00529-6.

[51] Badreddine Mandhouj, Mohamed Ali Cherni, and Mounir Sayadi. "An automated classification of EEG signals based on spectrogram and CNN for epilepsy diagnosis". In: *Analog integrated circuits and signal processing* 108.1 (2021), pp. 101–110.

[52] Dominique Makowski et al. "NeuroKit2: A Python toolbox for neurophysiological signal processing". In: *Behavior Research Methods* 53.4 (2021), pp. 1689–1696. DOI: 10.3758/s13428-020-01516-y. URL: https://doi.org/10.3758%2Fs13428-020-01516-y.

[53] Mark Claypool et al. "Surrender at 20? Matchmaking in league of legends". In: *2015 IEEE Games Entertainment Media Conference (GEM)*. IEEE. 2015, pp. 1–4.

[54] Sohye Lim and Byron Reeves. "Computer agents versus avatars: Responses to interactive game characters controlled by a computer or other player". In: *International Journal of Human-Computer Studies* 68.1-2 (2010), pp. 57–68.

[55] Amin Noroozi et al. "An efficient machine learning approach for extracting eSports players' distinguishing features and classifying their skill levels using symbolic transfer entropy and consensus nested cross-validation". In: *International Journal of Data Science and Analytics* (2024), pp. 1–14.

[56] Yu-Xuan Yang et al. "A recurrence quantification analysis-based channel-frequency convolutional neural network for emotion recognition from EEG". In: *Chaos: an interdisciplinary journal of nonlinear science* 28.8 (2018).

[57] Ateke Goshvarpour, Ataollah Abbasi, and Atefeh Goshvarpour. "Recurrence quantification analysis and neural networks for emotional EEG classification". In: *Applied Medical Informatics* 38.1 (2016), pp. 13–24.

[58] Andrew Webb et al. "To ensemble or not ensemble: When does end-to-end training fail?" In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III*. Springer. 2021, pp. 109–123.

[59] Pouya Bashivan et al. "Learning representations from EEG with deep recurrent-convolutional neural networks". In: *arXiv preprint arXiv:1511.06448* (2015).

[60] Onur Erdem Korkmaz, Sevde Gül Korkmaz, and Onder Aydemir. "Detection of multitask mental workload using gamma band power features". In: *Neural Computing and Applications* 36.18 (2024), pp. 10915–10926.

[61] Monicque M Lorist et al. "The influence of mental fatigue and motivation on neural network dynamics; an EEG coherence study". In: *Brain research* 1270 (2009), pp. 95–106.

[62] A Pokryszko-Dragan et al. "Stimulated peripheral production of interferon-gamma is related to fatigue and depression in multiple sclerosis". In: *Clinical neurology and neurosurgery* 114.8 (2012), pp. 1153–1158.

[63] Darshil Shah, Gopika Gopan K, and Neelam Sinha. "An investigation of the multidimensional (1D vs. 2D vs. 3D) analyses of EEG signals using traditional methods and deep learning-based methods". In: *Frontiers in Signal Processing* 2 (2022), p. 936790.

[64] Georgios N Dimitrakopoulos et al. "Task-independent mental workload classification based upon common multiband EEG cortical connectivity". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.11 (2017), pp. 1940–1949.

[65] Mahsa Bagheri and Sarah D Power. "Simultaneous classification of both mental workload and stress level suitable for an online passive brain–computer interface". In: *Sensors* 22.2 (2022), p. 535.

[66] Minkyu Ahn and Sung Chan Jun. "Performance variation in motor imagery brain–computer interface: a brief review". In: *Journal of neuroscience methods* 243 (2015), pp. 103–110.

[67] Ioannis Bikas et al. "Mental wear and tear: An exploratory study on mental fatigue in video games using the example of league of legends". In: *International Conference on Entertainment Computing*. Springer. 2022, pp. 125–139.

[68] Steffen Bickel, Michael Brückner, and Tobias Scheffer. "Discriminative learning under covariate shift." In: *Journal of Machine Learning Research* 10.9 (2009).

A Appendix

Match	Player	Type	EEG	GSR	HR
match_0	player_1	bot	15.95	15.95	15.97
match_1	player_1	real	25.53	25.53	25.55
match_2	player_1	bot	17.87	17.87	17.88
match_3	player_1	real	26.53	26.55	0.40
match_8	player_3	real	25.38	25.38	25.40
match_8	player_4	real	25.38	25.38	25.40
match_9	player_3	bot	16.98	16.98	17.00
match_9	player_4	bot	16.98	16.98	17.00
match_10	player_3	real	4.43	22.83	22.85
match_10	player_4	real	3.93	22.83	22.85
match_11	player_1	real	34.57	38.58	38.58
match_11	player_2	real	34.57	38.57	38.58
match_11	player_3	real	38.57	38.57	38.58
match_11	player_4	real	34.57	38.57	38.58
match_12	player_1	bot	11.25	15.12	15.13
match_12	player_2	bot	11.25	15.12	15.13
match_12	player_3	bot	15.12	15.12	15.13
match_12	player_4	bot	11.25	15.12	15.13
match_13	player_1	real	32.62	34.85	34.87
match_13	player_2	real	32.62	34.85	33.70
match_13	player_3	real	34.85	34.85	34.87
match_13	player_4	real	32.62	34.85	34.87
match_14	player_1	real	25.20	32.92	32.93
match_14	player_2	real	25.20	32.92	32.93
match_14	player_4	real	24.68	32.92	32.93
match_18	player_1	real	39.03	40.00	36.00
match_18	player_2	real	39.03	40.00	40.02
match_19	player_1	bot	17.38	17.38	17.40
match_19	player_2	bot	17.38	17.38	17.40
match_20	player_1	real	23.50	23.50	23.52
match_20	player_2	real	23.50	23.50	23.52
match_21	player_1	bot	16.98	16.98	17.00
match_21	player_2	bot	16.98	16.98	17.00
match_21	player_3	bot	16.98	16.98	17.00

Table 8: Details of the matches used in the analysis. Values indicate the length of the recorded signal in minutes.