Generalized Arousal Prediction through Machine Learning

Shaila Zaman, Panagiotis Tsiamytzis, and Ioannis Pavlidis, *Senior Member, IEEE*

Abstract—The abstract goes here.

Index Terms—Arousal, perinasal perspiration, affective computing.

1 Introduction

PSYCHOPHYSIOLOGICAL arousal describes the state of feeling awake, activated, and highly reactive to stimuli; it is important in regulating consciousness, attention, alertness, and information processing. Arousal is associated with fight or flight responses and linked to emotional states [1]. Per the Yerkes-Dodson law, an optimal level of arousal for performance exists, and too little or too much arousal can adversely affect task performance [2]. For this and other reasons, arousal has attracted considerable attention in human-machine interaction research.

In affective computing, there are two major schools of thought when it comes to arousal research - observational and physiological. The observational school of thought has been studying arousal in the context of the valence-arousal model of emotions. Relevant works focused on estimating ratings for valence and arousal through computer vision analysis of facial expressions. These estimates were validated against ground-truth provided by expert coders in the two-dimensional emotion space [3], [4]. As the image and video datasets used in the said investigations were assembled from the Internet, other ground-truthing options were limited. In some cases, valence-arousal predictions were aided by audio cues [5]. Several Machine Learning (ML) and Deep Learning (DL) algorithms were used to operationalize valence-arousal methods, including Support Vector Machines (SVM) [4], Convolutional Neural Networks (CNN) [4], and Long-Short Memory (LSTM) networks [3].

The physiological school of thought has been studying arousal through physiological signals obtained via imaging or wearable sensors. As these signals are acquired in controlled experiments, physiological arousal has been ground-truthed via self-assessment from participants. Electroencephalography (EE) signals have shown promise as arousal-valence estimators [6]. Some studies highlighted the potential of electrodermal activity (EDA) in measuring arousal [7]. Cardiovascular and respiratory indicators, including heart rate [8], heart rate variability, [9], and breathing [10] have also been found to associate with arousal responses. Researchers have also reported multimodal physiological methods for the determination of arousal. Recent work has been attempting to bring together the observational and physiological schools of thought by fusing facial observations and physiological signals into an integrated arousal prediction framework [11].

Martinez et al. [12] attempted to move physiological research on arousal to a more naturalistic setting, by monitoring 657 participants over the course of eight weeks. The monitoring focused on the recording of HRV from a wearable sensor and was coming complete with ecological momentary assessments (EMA) assessing perceived stress. The researchers found that HRV did not explain more than 2.2% of the variance of EMA responses. This result demonstrated that HRV may be a reliable measure of perceived stress during stressful tasks in the laboratory, but its reliability in naturalistic studies is in question. Furthermore, this result casts doubt on the validity of self-reported stress ratings as ground-truth for physiological indicators of arousal. Such self-reported stress ratings may be reflecting memory biases, coping responses [13], or chronic stressors that may not necessarily influence cardiovascular indicators.

2 Research Aims

The purpose of our research is twofold: First, to establish a universal model of arousal that traverses physiological modalities and application domains. Second, to establish a method for arousal rating that is objective and continuous.

Arousal is a physiological mechanism that helps humans to cope with demands and challenges imposed upon them by daily life. The sympathetic system does not have access to a different set of resources for producing arousals, say, during driving versus report writing. Hence, any type of physiological sensor in any domain of human endeavor picks up manifestations of the same underlying phenomenon. We posit that if people's physiological signals from any sensor over any task are properly curated and normalized to reduce noise and inter-individual reliability, their distributions must look very similar. Assuming such a universal property exists, then any arousal, irrespective of the task domain and recording sensor, can be described by a single parametric function. The latter suggests that if such a universal distribution is rated once, in a convenient study

• *S. Zaman and I. Pavlidis are with the Computational Physiology Laboratory, University of Houston, Houston TX.*
E-mail: ipavlidis@uh.edu

Manuscript received January 15, 2023.

with solid ground-truth, this rating can be carried out to any other study for which reliable ground-truth is not possible.

The remaining question is what the method of arousal rating would be. To overcome the limitations of present arousal rating methods, the new method must be objective and not subjective, and must have temporal resolution that matches the continuous nature of physiological signals. In other words, the new arousal rating method must be rooted in psychophysiological theory, rather than attempting to bridge conscious perceptions with a subconscious process, which arousal is.

3 Comparison to Related Work

Global vs. Domain Specific Application.
Any Sensor vs. Specific Sensors.
Limited Need for Ground-truthing.
Unlimited Potential for Transfer Learning.

4 Studies

4.1 Driving Simulator Study 1 or S1

The aim of the S1 study was to deconstruct the sources of distractions while driving and investigate separately their effects. To fulfill this objective, the study design included physical, cognitive, and emotional stressors. The study was carried out in a high fidelity simulator, where the participants had to drive the same itinerary several times. The itinerary was a simple drive with traffic in the opposing lanes but no traffic in the participants' own lane. The drive was 10.8 km long and consisted of five phases, from P_1 to P_5; in phases P_2 (3.2 km) and $P4$ (3.2 km) the stressor was applied, while phases P_1 (1.2 km), P_3 (2.8 km), and P_5 (0.4 km) were stressor free. The drive itself was designed to be so simple, so that cannot stimulate arousal by itself. The only possibility for arousal rested with the exogenous stressors in phases P_2 and P_4.

Each time a participant was driving the itinerary, a different stressor was applied (physical, cognitive, or emotional), and the order of application was randomized to ameliorate order effects. Prior to driving, participants underwent a baseline session, where they were measured for 5 minutes while listening to soothing music in a dimly lit room. The purpose of this session was to help in the estimation of the participants' homeostatic values.

The physical stressor was operationalized as texting while driving, where participants had to text back nondescript words, sent one by one to their smartphone. The cognitive stressor was operationalized as mathematical and analytical questions. The emotional stressor was operationalized by asking participants emotionally stirring questions. Our dataset includes data from $n = 31$ S1 participants (21 females/10 males).

4.2 Driving Simulator Study 2 or S2

The aim of the S2 study was to test if biofeedback can lead to prompt disengagement from physical and cognitive stressors while driving. For that reason, the experiment featured two arms: a non-interventional arm (without biofeedback), and an interventional arm (with biofeedback). Here we are concerned only with the non-interventional arm of the experiment, where the application of the stressful stimuli remained intact during the planned periods, much like in study S1. Study S2, however, had two main major differences from study S1: First, it lacked an emotional stressor, and second, featured a milder cognitive stressor, where participants had to subtract 13 from 1,022 instead of answering challenging math and analytical questions. Our dataset includes data of $n = 21$ S2 participants (10 females/11 males).

4.3 Driving Test Track Study 1 or T1

T1 was a driving study with an actual vehicle that took place in a test track facility in College Station, Texas. The aim of T1 was to implement an S1 type of design in naturalistic conditions. Each itinerary consisted of two runs up and down of an old airport runway, involving four U-turns. The total length of the itinerary was about 6.5 km. The stressors were applied on the driver during the straight segments of the drive. T1 featured physical and cognitive stressors. The cognitive stressor was like that of S2. The order of physically and cognitively distracted drives was randomized to ameliorate order effects. Our dataset includes data from $n = 19$ T1 participants (11 females/8 males).

4.4 Office Task Study or OT

OT was a controlled experiment, aiming to investigate arousal in different knowledge work tasks. A key task examined in OT was essay writing. Participants were given 50 min to compose an essay on the topic of technological singularity, that is when machines overtake human intelligence. The aim of this task was to simulate prototypical knowledge work that induces mental load, interspersed with email distractions that may contribute to arousal. Our dataset includes data of OT's essay task for $n = 44$ participants (29 females/15 males).

4.5 Deadline Study or DS

DS was a naturalistic study aiming to investigate arousal patterns of academic researchers around deadlines. The protocol included four days of observation - two days leading to the participants' deadline and two days following it. During the monitoring days the participants were asked to work at the office, as they would normally do. No restrictions were applied to their movements and activities. Our dataset includes data from the moments $n = 8$ participants were working at their office computers during the said days.

5 Methods

5.1 Quality Control of Physiological Data

The S1 and OT dataset descriptors have been published in *Scientific Data*, and their data underwent rigorous quality control and validity checks [14], [15]. Hence, for these two studies, we use the curated data given in the respective Open Science Framework (OSF) repositories [16], [17]. For the S2, T1, and DS studies, we curated the data by following similar quality control methods reported in S1 and OT. We removed heart rate and breathing rate signal values outside

$[40, 140]$ BPM and $[4, 40]$ BPM, respectively. Healthy sitting subjects under low to moderate stressors are unlikely to have heart and breathing rate values beyond the said ranges.

For perinasal perspiration (PP), signal extraction from thermophysiological imagery is morphological in nature, quantifying the presence of active perspiration pores. These pores are tiny in size, and thus PP signal quality depends on sharp focusing. Focus quality in facial thermal imaging sequences is assessed on the basis of edge contrast between the participants' (cold) eyebrows and (hot) surrounding tissue. We did not find any improperly focused thermal imagery in S2, T1, and DS. PP signal quality also depends on the performance of the tissue tracker, which locks on the participants' perinasal area while they naturally move their head as they drive or do computer work. Momentary tracking failures manifest as spikes in the PP signals, which can be easily detected and removed.

5.2 Physiological Signal Distributions and Arousal Universality

People have different homeostatic points, which manifest as different baseline values in their PP, HR, and BR data. To ameliorate these interindividual differences and allow a less biased manifestation of arousal to come to the fore, we normalize the participants' physiological values by subtracting their estimated baseline levels:

$$\Delta PP_i(t) = \ln PP_i(t) - \overline{\ln PP}_{BL_i} \quad (1)$$

$$\Delta HR_i(t) = HR_i(t) - \overline{HR}_{BL_i} \quad (2)$$

$$\Delta BR_i(t) = BR_i(t) - \overline{BR}_{BL_i} \quad (3)$$

where $PP_i(t)$, $HR_i(t)$, and $BR_i(t)$ are the perinasal perspiration, heart rate, and breathing rate of participant i at observation time t, while $\overline{PP}_{BL_i}$, $\overline{HR}_{BL_i}$, and $\overline{BR}_{BL_i}$ are the mean perinasal perspiration, heart rate, and breathing rate values of participant i during the baseline period of observation. Among the three physiological variables, PP is known to feature more skewed distributions with respect to the other two [18]. For this reason, and per standard literature practice [15], we log-transform the perinasal perspiration values to bring their skewness closer to that of the HR and BR measures.

Figure 1 shows the normalized physiological distributions and the corresponding Q-Q plots for all studies and their combinations. Ideally, zero would indicate the point of homeostatic control. This may not be exactly the case, however, because of imperfections in the baselining process. The distribution's mean is another point estimate of homeostatic control, which falls close to the zero point. Common to all studies, the area enveloping the two point estimates of homeostatic control (i.e., zero and mean) lies in the normal region of the distribution. In contradistinction, the tails of the distributions have patterned differences. In study $\mathcal{S}1$, we observe that the PP and HR distributions are right-skewed, while the BR distribution has fat tails. The right-skeweness of PP and HR suggests upregulation of high-arousal, and is in agreement with reports in psychological studies [19]. The fat tails in the BR distribution, however, are surprising, suggesting both upregualtion of high-arousal and downregulation of low-arousal. We posit that the downregulation of low-arousal in the BR distribution is an artifact of breathing modulation due to speech. In the cognitively and emotionally distracted drives of study $\mathcal{S}1$, the drivers were speaking their answers to the questions they were receiving [20]. To accommodate speech production, breathing volume increases while breathing frequency decreases [10]. In study $\mathcal{S}2$, the PP and HR distributions remain right-skewed, but the fat tails in the BR distribution have been reduced, because the $\mathcal{S}2$ dataset includes one less 'speaking drive' [21]. The same conditions and observations apply in study $\mathcal{R}$. In study $\mathcal{O}$, the fat tails have totally receded, giving rise to a right-skewed breathing rate distribution that is similar to the distributions of PP and HR. We posit that this happens because in study $\mathcal{O}$ the participants write an essay in silence [15], and thus breathing rate is undisturbed by speech production.

5.3 Data Labeling

We choose the distributional means as more realistic point estimates of homeostasis, falling always close to the ideal zero points. Our methodological thesis is that normalized physiological measures of arousal have a Normal distribution in the vicinity of the homeostatic point, while they are right-skewed in the high-arousal region. Hence, arousal has universal nature, with an intermediate-arousal region standing between the low-arousal region to the left and the overextended high-arousal region to the right. As we have seen in Fig. 1, the normalized PP and HR distributions conform to this pattern across studies. The normalized BR distribution, however, conforms to the said universal pattern only when speech is absent. If speech is present in a study, there is moderate departure from the universal pattern, where not only the high-arousal but also the low-arousal region of BR becomes overextended (fat tails).

Assuming a universal arousal pattern, grading of arousal in one study distribution, would be applicable to any other study distribution. The key objective then is to find a study for which grading of physiological arousal can be done with confidence; that is, a study bearing reliable ground-truth. We chose to perform arousal grading on the physiological distributions of study $\mathcal{S}1$. Two reasons led us to this choice: First, there is strong evidence that $\mathcal{S}1$'s controlled stimuli did produce significant arousal on participants. This evidence is based on participants' recorded micro-tremors during the application of the stimuli [20] - a definitive sign of 'fight or flight' responses [22]. Thus, $\mathcal{S}1$ bears solid ground-truth information. Second, the $\mathcal{S}1$ study features a comprehensive set of arousal stimuli, including physical, cognitive, and emotional stressors [14]. Accordingly, the $\mathcal{S}1$ dataset is representative of a broad variety of naturally occurring arousal stimuli.

Arousal grading in a physiological distribution is tantamount to identifying the borders of arousal's intermediate state - to the left of it there would be low-arousal, while to the right of it there would be high-arousal. Such a distributional division is in congruence with the famous Yerkes-Dodson law. Our initial guess was that the intermediate state of arousal would likely be delineated by one standard deviation from the mean (i.e., homeostatic point) in $\mathcal{S}1$'s physiological distributions, that is, $\mu_k \pm \mathrm{sd}_k$, where $k \equiv PP_N$ or HR_N or BR_N. Figure 2a shows how

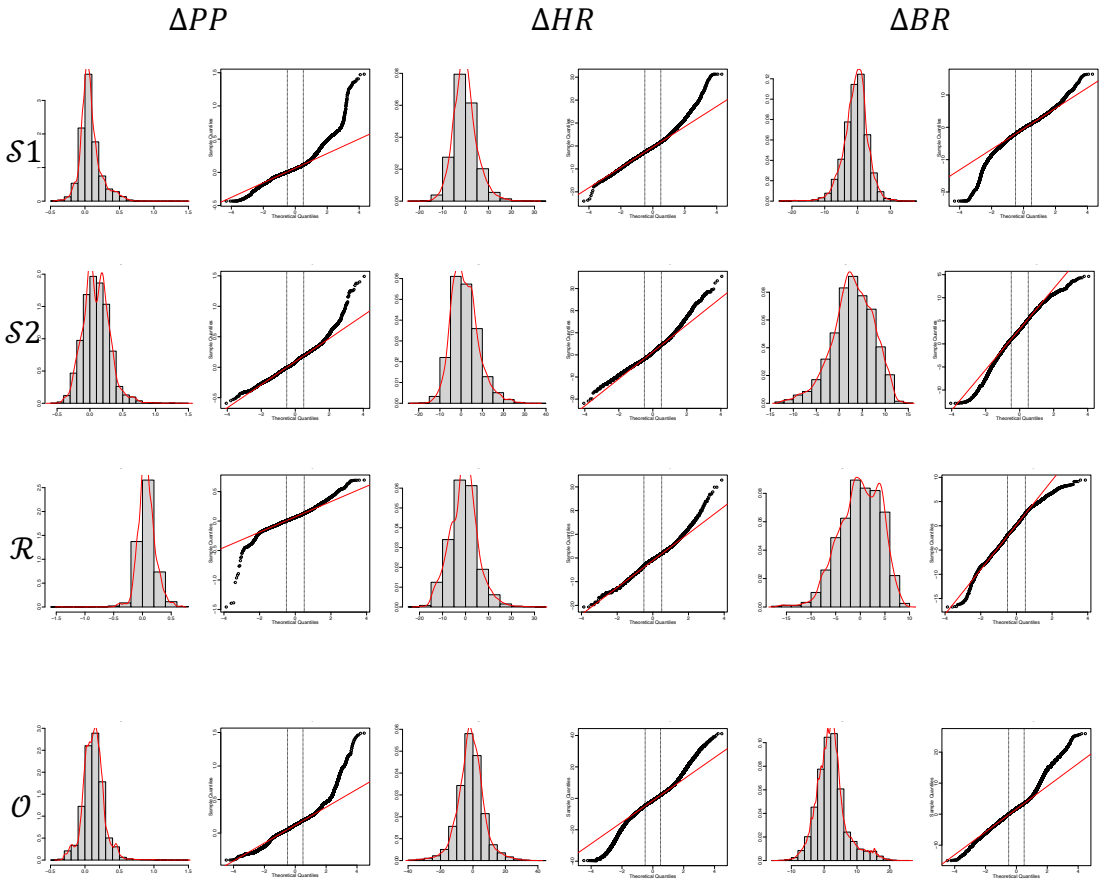

Fig. 1: **Normalized physiological distributions for all studies and combinations thereof.**

$\mathcal{S}1$'s experimental timeline appears when we apply such a threshold. In $\mathcal{S}1$ there were two periods of stimulus application, where we know that the majority of participants exhibited micro-tremors, and thus significant arousal [20]. We also know that soon after the removal of the stimulus, micro-tremors died down [20], which suggests that the participants returned to a state of low-arousal. This was the expressed purpose of $\mathcal{S}1$'s experimental design, that is, to interlace periods of trivial driving in a straight road and with no traffic, with challenging periods punctuated by strong distractions. The labeling in Fig. 2a does not capture the said situation well. It shows few high-arousal states during the challenging periods, where we would expect far more, while it shows a preponderance of intermediate-arousal, even in periods we know micro-tremors not only receded but disappeared altogether.

This outcome suggested that we needed to tighten our threshold, and accordingly we chose the intermediate state of arousal to be delineated by half a standard deviation from the mean in $\mathcal{S}1$'s physiological distributions, that is, $\mu_k \pm 0.5\text{sd}_k$, where $k \equiv PP_N$ or HR_N or BR_N. Figure 2c shows how $\mathcal{S}1$'s experimental timeline appears when we apply this tighter threshold. The new labeling tends to capture well $\mathcal{S}1$'s experimental design across all physiological channels. It shows for most participants two main periods of arousal, corresponding to the two times the stressors were applied. The arousal periods are usually enveloped by periods of low-arousal, corresponding to absence of micro-tremors and matching the triviality of the non-distracted driving task. Intermediate arousal plays a more limited role, and it appears to act either as a transition between high-arousal and low-arousal, or as an alternative, when these two extreme states are not reached.

6 Analytic Results

We pursue two different analytic approaches aiming to investigate whether temporal context plays a role or not in arousal classification:

Time Independent Classification/TIC: In the time independent approach, the models have to classify 10 s data segments in the testing set, based on training they received on 10 s data segments in the training set. The data segments are considered temporally unrelated.

Time Dependent Classification/TDC: In the time dependent approach, the models have to classify in the testing set 10 s data segments into the future by looking into 30 s of their immediate past. TDC draws on training received on pairs of 10 s data segments preceded by 30 s data segments in the training set. The rolling window to operationalize TDC is 5 s.

In both TIC and TDC, we report classification results for different training designs. These designs correspond to the

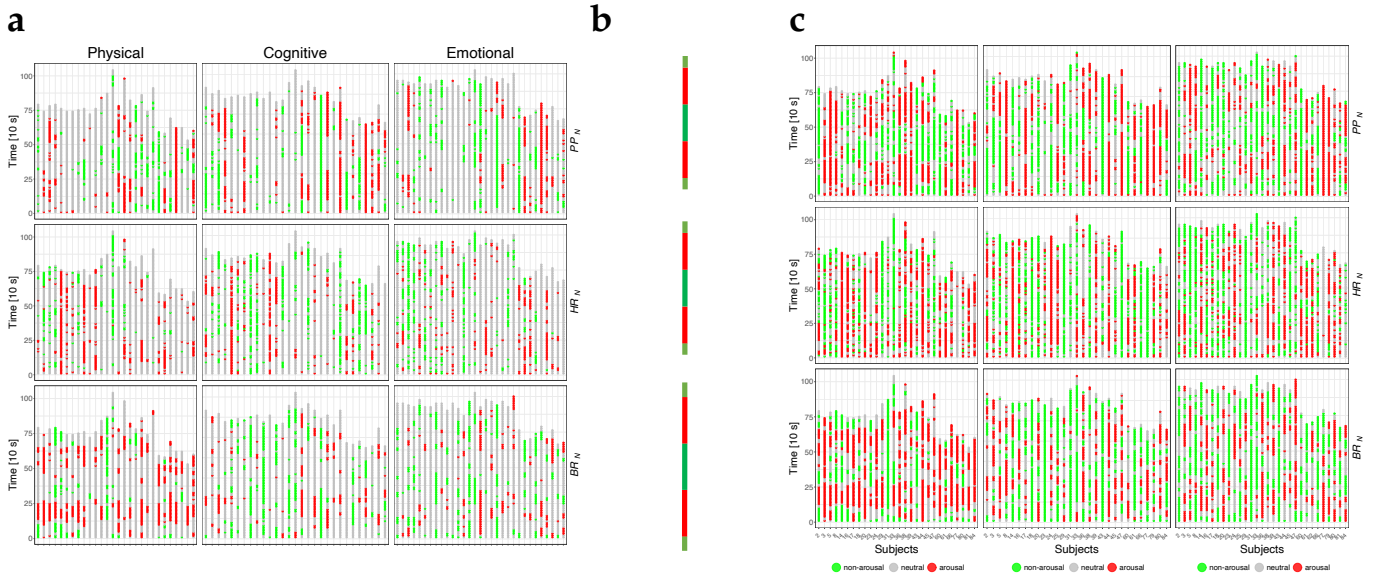

Fig. 2: **State of arousal for each 10 s window in the $\mathcal{S}1$ study.** Panel rows correspond to different physiological channels, with PP_N, HR_N, and BR_N indicating normalized perinasal perspiration, heart rate, and breathing rate, respectively. The horizontal axes of the panels are pegged with the $n = 31$ usable subjects of $\mathcal{S}1$. The colored timelines in the middle of the figure depict the ideal succession of low-arousal and high-arousal periods aimed by the experimental design, serving as a comparative yardstick. **a.** Intermediate arousal state is delineated by one standard deviation from the distributional means. **b.** Intermediate arousal state is delineated by half standard deviation from the distributional means.

following training study combinations: single study vs. multiple controlled driving studies vs. multiple studies across the realism axis vs. multiple studies across the domain axis. The results are presented in matrices, where the rows x_i correspond to groups of training studies, whereas the columns y_j correspond to groups of testing studies. A study group may contain multiple or, trivially, a single study. Each cell in these matrices shows the Area Under the Curve (AUC) of classifying in study group y_j, after training in study group x_i. We do not report AUC values for the diagonal cells of matrices, which correspond to identical training-testing pairs. Results from such trivial pairs are of little interest to our universality-oriented arousal research. For each training design, we report results per physiological channel, that is, PP, HR, or BR. We also report results per combination of physiological channels, that is, PP_BR, PP_HR, HR_BR, or PP_HR_BR.

6.1 Modeling of Performance Comparison

We construct Eq. (4) to quantify the effect of Machine Learning (ML) and physiological ground-truthing (GT) methods on the AUC performance of testing-training study pairs.

$$\begin{aligned} AUC_k(i,j) \sim & \beta_0 + \beta_1 ML(i) + \beta_2 GT(j) + \\ & \beta_3 M(i) \times GTS(j) + 1|C_k. \end{aligned} \tag{4}$$

Equation (4) represents a family of four models, where $k \in \{SS, MS, RI, DI\}$, that is, the response variable AUC_k is examined in four different combinatorial scenarios:

1) **Single-Study scenario $k = SS$:** The training-testing study pairs are constituted solely by single study datasets.
2) **Multi-Study scenario $k = MS$:** It includes a dataset amalgamating more than one controlled studies. This is the dataset S12 = S1 ∪ S2 that encompasses both the S1 and S2 driving simulator studies.
3) **Realism Integration scenario $k = RI$:** It includes a dataset amalgamating controlled and naturalistic studies in the same domain. This is the dataset S12R = S1 ∪ S2 ∪ R that encompasses the driving simulator studies S1 and S2, as well as the racetrack study R.
4) **Domain Integration scenario $k = DI$:** It includes a dataset amalgamating controlled studies from different domains. This is the dataset S12O = S1 ∪ S2 ∪ O that encompasses the driving simulator studies S1 and S2, as well as the office tasks study O.

In Eq. (4), $ML(i)$ is a categorical variable with two levels $i \in \{RF, LSTM\}$, corresponding to the Random Forest and Long-Short Term Memory methods employed in this research as examples of time-independent and time-depended classification, respectively. We consider the level RF as the base level of variable ML. $GT(j)$ is a categorical variable with seven levels $j \in \{PP, HR, BR, PP\_BR, PP\_HR, HR\_BR, PP\_HR\_BR\}$, corresponding to the various ways of performing physiological ground-truthing with the PP, HR, and BR signals or combinations thereof. We consider the level BR as the base level of variable GT. The term $ML(i) \times GT(j)$ denotes the interaction between the categorical variables $ML(i)$ and $GT(j)$. As each training-testing combination (cell) is treated with different ML and physiological ground-truthing methods, this is a repeat-measures design. Accordingly, in Eq. (4) we use

random-centered effects $1|C_k$ on the training-testing cells of each scenario k.

6.2 Time Independent Classification Results

Training with Single Study. The combinatorial matrix in Fig. 3a shows that the HR channel attains the best performance across the board with cross-study AUC $\in [0.85, 0.98]$, while the BR channel fares the worst with cross-study AUC $\in [0.64, 0.95]$. The BR performance is especially poor when the controlled driving study S2 is involved either in training or testing. The PP channel performs on par with the HR channel, unless the controlled knowledge work study OT is used either for training or for testing; then, the PP performance drops. When data from more than one physiological channels are used in training, then the combinations that include BR show marked performance improvement with respect to the sole BR channel; e.g., for PP_BR cross-study AUC $\in [0.83, 0.97]$.

Receiver Operating Characteristic (ROC) Curves. Figure 4 shows the best and worst ROC performing training-testing pairs for each sensor channel. We observe that S1 is always present in the best performing pairs, while S2 is almost always present in the worst performing pairs. This explains the AUC snapshots in Fig. 3a, bringing to the fore that quality of controlled studies matters when it comes to training potency. S2 was lacking emotional stimuli and had weaker cognitive stimuli with respect to S1. While in real life sources of arousal vary, in controlled experiments, this is not always the case. For a controlled study dataset to be effective in universal arousal training, it is important to feature a comprehensive mix of arousal sources. Furthermore, as compensation to their artificial character, controlled studies should be designed with strong stimuli, when universal training potency is a key objective.

Training with Multiple Controlled Driving Studies. In the combinatorial matrix in Fig. 3b, SS = S1 $\cup$ S2, that is, a dataset encompassing more than one controlled driving study. It is evident that when multiple controlled driving studies are used in training, then the performance of the BR channel markedly improves with respect to single study training, exhibiting now AUC $\in [0.82, 0.98]$. This improvement reaches perfection when a combination of BR and other physiological channels partake in training; e.g., for HR_BR the cross-study AUC $\in [0.94, 0.97]$.

Training with Driving Studies Across the Realism Axis. In the combinatorial matrix in Fig. 3c, SST = S1 $\cup$ S2 $\cup$ T1, that is, a dataset encompassing both controlled and naturalistic driving studies. It is evident that when driving studies across the realism axis are used in training, then the bottom performance of the BR channel markedly improves with respect to single study training, exhibiting now cross-study AUC $\in [0.83, 0.85]$.

Training with Studies Across the Domain Axis. In the combinatorial matrix in Fig. 3d, SSO = S1 $\cup$ S2 $\cup$ OT, that is, a dataset encompassing both driving and knowledge work studies. It is evident that when studies from different domains are used in training, then the performance of all physiological channels nears perfection. Specifically, the PP channel has cross-study AUC $\in [0.92, 0.95]$, the HR channel has cross-study AUC $\in [0.97, 0.98]$, and the BR channel has cross-study AUC $\in [0.92, 0.97]$.

6.3 Time Dependent Classification Results

Training with Single Study. The combinatorial matrix in Fig. 5a shows that in TDC the role of the HR and BR channels has been reversed with respect to TIC. The BR channel now attains the best performance with cross-study AUC $\in [0.79, 0.97]$, while the HR channel fares the worst with cross-study AUC $\in [0.80, 0.87]$. The BR performance is especially poor when the controlled driving study S2 is involved either in training or testing. The PP channel's performance is in the middle, as in TIC. The role of the specific contributors to this performance, however, has been reversed with respect to TIC. The PP performance peaks when the controlled knowledge work study OT is used either for training or for testing.

Training with Multiple Controlled Driving Studies. In the combinatorial matrix in Fig. 5b, SS = S1 $\cup$ S2, that is, a dataset encompassing more than one controlled driving study. In TDC, when multiple controlled driving studies are used in training, then the performance of the HR channel does not markedly improve with respect to single study training, exhibiting AUC $\in [0.82, 0.88]$; HR remains the bottom performer.

Training with Driving Studies Across the Realism Axis. In the combinatorial matrix in Fig. 5c, SST = S1 $\cup$ S2 $\cup$ T1, that is, a dataset encompassing both controlled and naturalistic driving studies. In TDC, when driving studies across the realism axis are used in training, then the performance of the HR channel does not markedly improve with respect to single study training, exhibiting cross-study AUC $\in [0.85, 0.86]$; HR continues to be the bottom performer.

Training with Studies Across the Domain Axis. In the combinatorial matrix in Fig. 5d, SSO = S1 $\cup$ S2 $\cup$ OT, that is, a dataset encompassing both driving and knowledge work studies. In TDC, when studies from different domains are used in training, then the performance of the HR channel catches up with the performance of the PP channel, but both remain below the performance of the BR channel. Specifically, the PP channel has cross-study AUC $\in [0.81, 0.87]$, the HR channel has cross-study AUC $\in [0.83, 0.90]$, and the BR channel has cross-study AUC $\in [0.93, 0.98]$.

7 Discussion

This will be the Discussion section.

Acknowledgments

The authors would like to thank... [15]

TABLE 1: **Results of the optimized mixed effects driving model described by Eq. (4)**. $\beta.$ stands for the coefficient estimates. Significance levels have been set as follows: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.

Predictor	$\beta.$	Standard Error	Degrees of Freedom	t value	Pr($>\|t\|$)	
SINGLE-STUDY SCENARIO WHERE THE RESPONSE VARIABLE IS AUC_{SS}						
$ML[LSTM]$	0.080	0.017	143.000	4.602	<0.001	***
$GT[PP]$	0.075	0.017	143.000	4.311	<0.001	***
$GT[HR]$	0.118	0.017	143.000	6.711	<0.001	***
$GT[PP\_HR]$	0.141	0.017	143.000	8.080	<0.001	***
$GT[PP\_BR]$	0.081	0.017	143.000	4.658	<0.001	***
$GT[HR\_BR]$	0.100	0.017	143.000	5.760	<0.001	***
$GT[PP\_HR\_BR]$	0.101	0.017	143.000	5.793	<0.001	***
$ML[LSTM] \times GT[PP]$	−0.103	0.025	143.000	−4.183	<0.001	***
$ML[LSTM] \times GT[HR]$	−0.163	0.025	143.000	−6.611	<0.001	***
$ML[LSTM] \times GT[PP\_HR]$	−0.173	0.025	143.000	−7.022	<0.001	***
$ML[LSTM] \times GT[PP\_BR]$	−0.089	0.025	143.000	−3.627	<0.001	***
$ML[LSTM] \times GT[HR\_BR]$	−0.110	0.025	143.000	−4.458	<0.001	***
$ML[LSTM] \times GT[PP\_HR\_BR]$	−0.103	0.025	143.000	−4.177	<0.001	***
MULTI-STUDY SCENARIO WHERE THE RESPONSE VARIABLE IS AUC_{MS}						
$ML[LSTM]$	0.057	0.022	65.000	2.551	0.013	*
$GT[PP]$	−0.001	0.022	65.000	−0.005	0.996	
$GT[HR]$	0.072	0.022	65.000	3.234	0.002	**
$GT[PP\_HR]$	0.075	0.022	65.000	3.373	0.001	**
$GT[PP\_BR]$	0.019	0.022	65.000	0.836	0.406	
$GT[HR\_BR]$	0.077	0.022	65.000	3.466	<0.001	***
$GT[PP\_HR\_BR]$	0.053	0.022	65.000	2.368	0.021	*
$ML[LSTM] \times GT[PP]$	−0.076	0.031	65.000	−2.427	0.018	*
$ML[LSTM] \times GT[HR]$	−0.161	0.031	65.000	−5.119	<0.001	***
$ML[LSTM] \times GT[PP\_HR]$	−0.156	0.031	65.000	−4.972	<0.001	***
$ML[LSTM] \times GT[PP\_BR]$	−0.072	0.031	65.000	−2.290	0.025	*
$ML[LSTM] \times GT[HR\_BR]$	−0.112	0.031	65.000	−3.578	<0.001	***
$ML[LSTM] \times GT[PP\_HR\_BR]$	−0.092	0.031	65.000	−2.918	0.005	**
REALISM-INTEGRATION SCENARIO WHERE THE RESPONSE VARIABLE IS AUC_{RI}						
$ML[LSTM]$	0.106	0.018	13.000	6.015	<0.001	***
$GT[PP]$	0.020	0.018	13.000	1.143	0.274	
$GT[HR]$	0.100	0.018	13.000	5.461	<0.001	***
$GT[PP\_HR]$	0.121	0.018	13.000	6.906	<0.001	***
$GT[PP\_BR]$	0.040	0.018	13.000	2.302	0.039	*
$GT[HR\_BR]$	0.119	0.018	13.000	6.764	<0.001	***
$GT[PP\_HR\_BR]$	0.108	0.018	13.000	6.162	<0.001	***
$ML[LSTM] \times GT[PP]$	−0.034	0.025	13.000	−1.373	0.193	
$ML[LSTM] \times GT[HR]$	−0.182	0.025	13.000	−7.314	<0.001	***
$ML[LSTM] \times GT[PP\_HR]$	−0.164	0.025	13.000	−6.579	<0.001	***
$ML[LSTM] \times GT[PP\_BR]$	−0.053	0.025	13.000	−2.137	0.005	
$ML[LSTM] \times GT[HR\_BR]$	−0.139	0.025	13.000	−5.573	<0.001	***
$ML[LSTM] \times GT[PP\_HR\_BR]$	−0.123	0.025	13.000	−4.959	<0.001	***
DOMAIN-INTEGRATION SCENARIO WHERE THE RESPONSE VARIABLE IS AUC_{DI}						
$ML[LSTM]$	0.007	0.019	13.000	0.345	0.736	
$GT[PP]$	−0.012	0.019	13.000	−0.599	0.559	
$GT[HR]$	0.029	0.019	13.000	1.491	0.160	
$GT[PP\_HR]$	0.021	0.019	13.000	1.066	0.306	
$GT[PP\_BR]$	−0.004	0.019	13.000	−0.022	0.983	
$GT[HR\_BR]$	0.025	0.019	13.000	1.304	0.215	
$GT[PP\_HR\_BR]$	0.001	0.019	13.000	0.067	0.947	
$ML[LSTM] \times GT[PP]$	−0.106	0.027	13.000	−3.882	0.002	**
$ML[LSTM] \times GT[HR]$	−0.121	0.027	13.000	−4.436	<0.001	***
$ML[LSTM] \times GT[PP\_HR]$	−0.121	0.027	13.000	−4.414	<0.001	***
$ML[LSTM] \times GT[PP\_BR]$	−0.074	0.027	13.000	−2.720	0.018	*
$ML[LSTM] \times GT[HR\_BR]$	−0.077	0.027	13.000	−2.798	0.015	*
$ML[LSTM] \times GT[PP\_HR\_BR]$	−0.061	0.027	13.000	−2.214	0.045	*

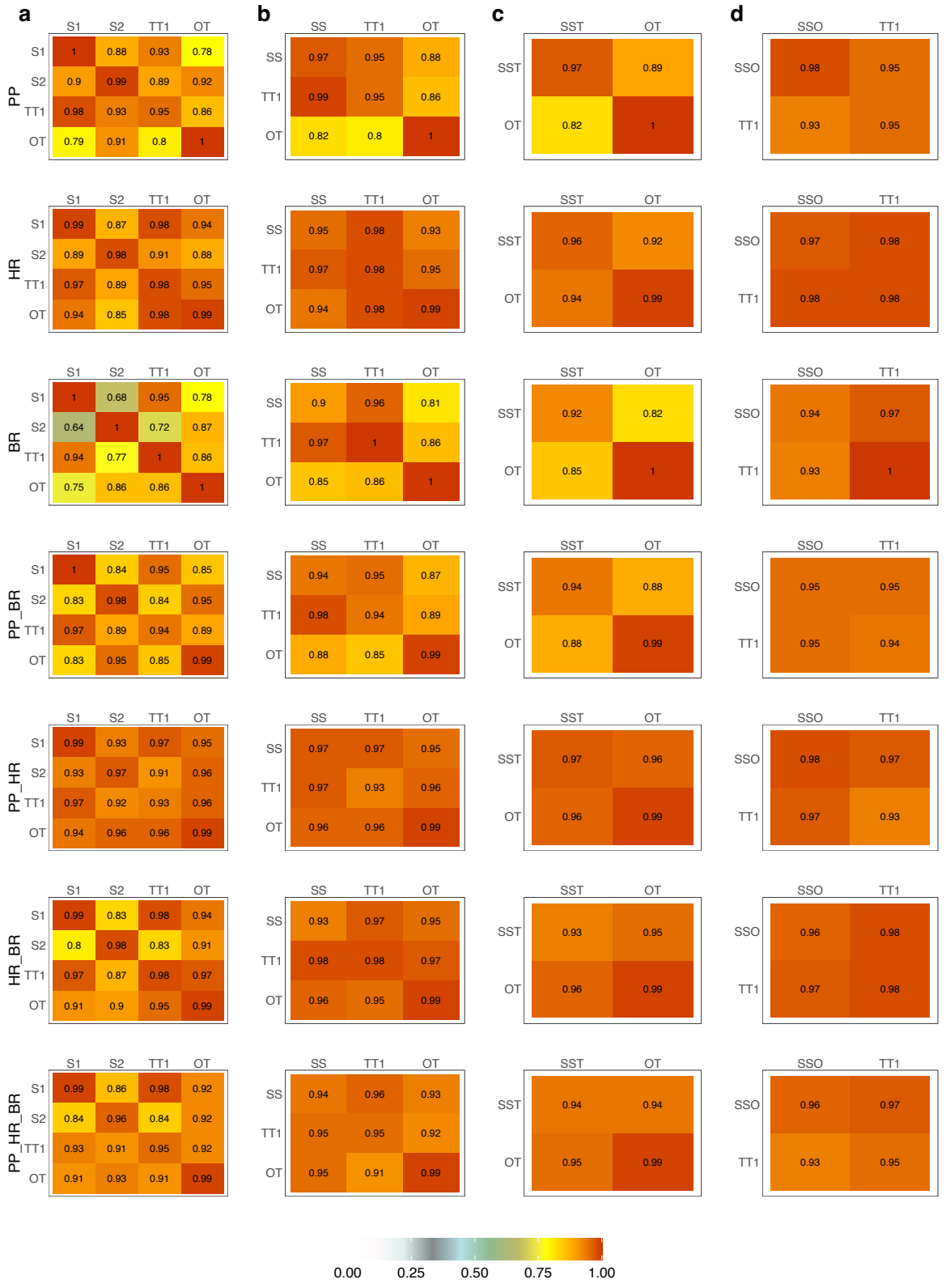

Fig. 3: **Time independent classification results in the form of combinatorial study matrices.** Matrix rows indicate training studies while matrix columns indicate testing studies. The symbols in the diagonal cells denote the studies that partake in the training-testing combinations. The numbers in the off-diagonal cells are the cross-study AUCs. Panel rows show classification results associated with sensing modalities PP, HR, BR and their combinations thereof. Panel columns show classification results associated with different study inclusion methods in the training sets. **Panels a.** Training sets include single studies. **Panels b.** Training sets include multiple driving controlled studies. **Panels c.** Training sets include driving studies across the realism axis. **Panels d.** Training sets include studies across the domain axis.

REFERENCES

[1] A. M. Herman, H. D. Critchley, and T. Duka, "The role of emotions and physiological arousal in modulating impulsive behaviour," *Biological psychology*, vol. 133, pp. 30–43, 2018.

[2] R. M. Yerkes, J. D. Dodson *et al.*, "The relation of strength of stimulus to rapidity of habit-formation," 1908.

[3] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, "AFEW-VA database for valence and arousal estimation in-the-wild," *Image and Vision Computing*, vol. 65, pp. 23–36, 2017.

[4] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[5] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[6] F. Galvão, S. M. Alarcão, and M. J. Fonseca, "Predicting exact valence and arousal values from eeg," *Sensors*, vol. 21, no. 10, p. 3414, 2021.

[7] T. Pakarinen, J. Pietilä, and H. Nieminen, "Prediction of self-perceived stress and arousal based on electrodermal activity," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 2191–2195.

[8] B. Reimer and B. Mehler, "The impact of cognitive workload on physiological arousal in young adult drivers: a field study and simulation validation," *Ergonomics*, vol. 54, no. 10, pp. 932–942, 2011.

[9] N. P. Murray, T. D. Raedeke *et al.*, "Heart rate variability as an indicator of pre-competitive arousal." *International Journal of Sport Psychology*, vol. 39, no. 4, pp. 346–355, 2008.

[10] S. A. Shea, "Behavioural and arousal-related influences on breathing in humans," *Experimental Physiology: Translation and Integration*, vol. 81, no. 1, pp. 1–26, 1996.

[11] L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E.-M. Messner, E. Cambria, G. Zhao, and B. W. Schuller, "The muse 2021 multimodal sentiment analysis challenge: sentiment, emotion, physiological-emotion, and stress," in *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, 2021, pp. 5–14.

[12] G. J. Martinez, T. Grover, S. M. Mattingly, G. Mark, S. D'Mello, T. Aledavood, F. Akbar, P. Robles-Granda, and A. Striegel, "Alignment between heart rate variability from fitness trackers and perceived stress: Perspectives from a large-scale in situ longitudinal study of information workers," *JMIR Human Factors*, vol. 9, no. 3, p. e33754, 2022.

[13] M. F. Scheier, J. K. Weintraub, and C. S. Carver, "Coping with stress: divergent strategies of optimists and pessimists." *Journal of personality and social psychology*, vol. 51, no. 6, p. 1257, 1986.

[14] S. Taamneh, P. Tsiamyrtzis, M. Dcosta, P. Buddharaju, A. Khatri, M. Manser, T. Ferris, R. Wunderlich, and I. Pavlidis, "A multimodal dataset for various forms of distracted driving," *Scientific data*, vol. 4, no. 1, pp. 1–21, 2017.

[15] S. Zaman, A. Wesley, D. R. D. C. Silva, P. Buddharaju, F. Akbar, G. Gao, G. Mark, R. Gutierrez-Osuna, and I. T. Pavlidis, "Stress and productivity patterns of interrupted, synergistic, and antagonistic office activities," *Scientific Data*, vol. 6, 2019.

[16] S. Taamneh, P. Tsiamyrtzis, M. Dcosta, P. Buddharaju, A. Khatri, M. Manser, T. Ferris, R. Wunderlich, and I. Pavlidis, "Simulator Study I – a multimodal dataset for various forms of distracted driving," *Open Science Framework* https://doi.org/10.17605/osf.io/c42cn, 2017.

[17] S. Zaman, A. Wesley, D. Cunha, P. Buddharaju, F. Akbar, G. Gao, G. Mark, R. Gutierrez-Osuna, and I. Pavlidis, "Office Tasks 2019 – a multimodal dataset," *Open Science Framework* https://doi.org/10.17605/osf.io/zd2tn, 2019.

[18] D. Shastri, A. Merla, P. Tsiamyrtzis, and I. Pavlidis, "Imaging facial signs of neurophysiological responses," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 2, pp. 477–484, 2009.

[19] S. V. Wass, K. Clackson, and V. Leong, "Increases in arousal are more long-lasting than decreases in arousal: On homeostatic failures during emotion regulation in infancy," *Infancy*, vol. 23, no. 5, pp. 628–649, 2018.

[20] I. Pavlidis, M. Dcosta, S. Taamneh, M. Manser, T. Ferris, R. Wunderlich, E. Akleman, and P. Tsiamyrtzis, "Dissecting driver behaviors under cognitive, emotional, sensorimotor, and mixed stressors," *Scientific Reports*, vol. 6, no. 1, p. 25651, 2016.

[21] I. Pavlidis, A. Khatri, P. Buddharaju, M. Manser, R. Wunderlich, E. Akleman, and P. Tsiamyrtzis, "Biofeedback arrests sympathetic and behavioral effects in distracted driving," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 453–465, 2021.

[22] P. E. Paredes, F. Ordonez, W. Ju, and J. A. Landay, "Fast & furious: detecting stress with a car steering wheel," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, 2018, pp. 1–12.

Shaila Zaman Biography text here.

Ioannis Pavlidis Dr. Pavlidis is the Eckhard-Pfeiffer Distinguished Professor of Computer Science and Director of the Computational Physiology Laboratory at the University of Houston. His research is funded by multiple sources including the National Science Foundation, transportation agencies, and medical institutions. He has published extensively in the areas of affective computing, data science, and science of science. He was the first to conceive and develop contact-free methods for measuring physiological variables, including electrodermal activity, breathing, and heart function, which he used to study stress in the wild.

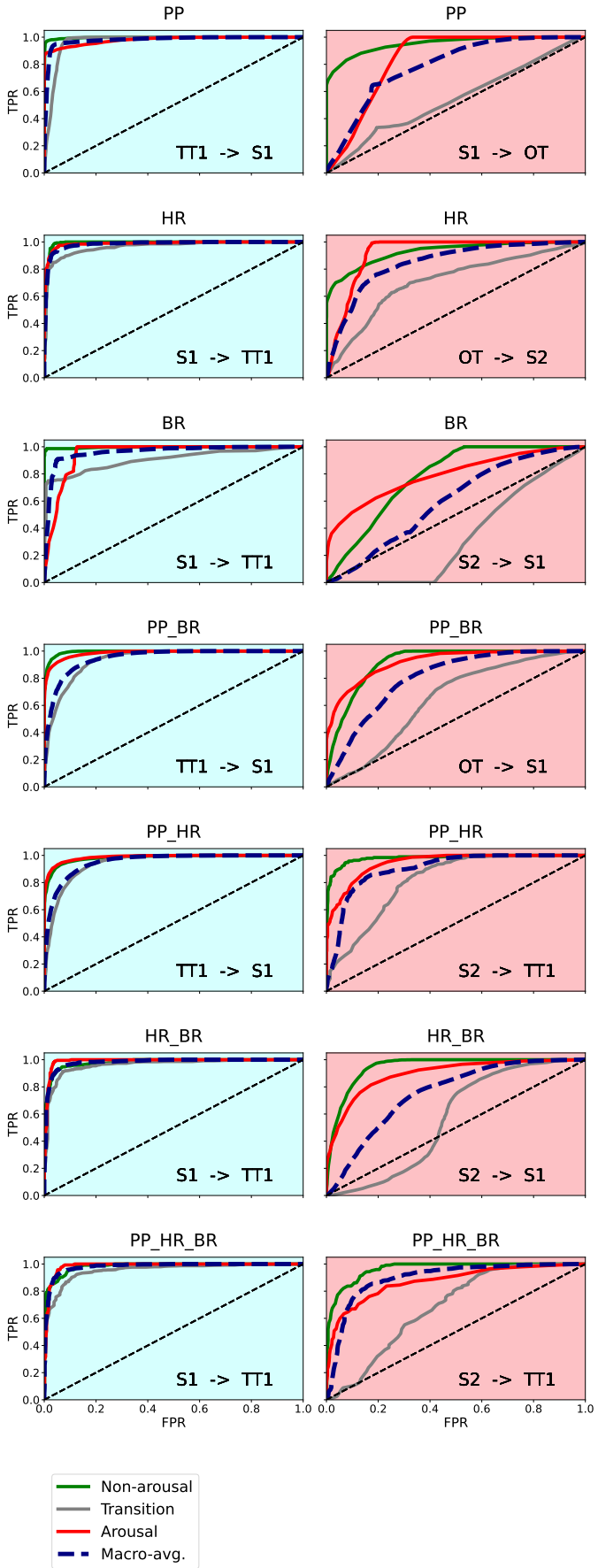

Fig. 4: **Best and worst Receiver Operating Characteristic (ROC) curves when training data are constituted with different sensor combinations.**

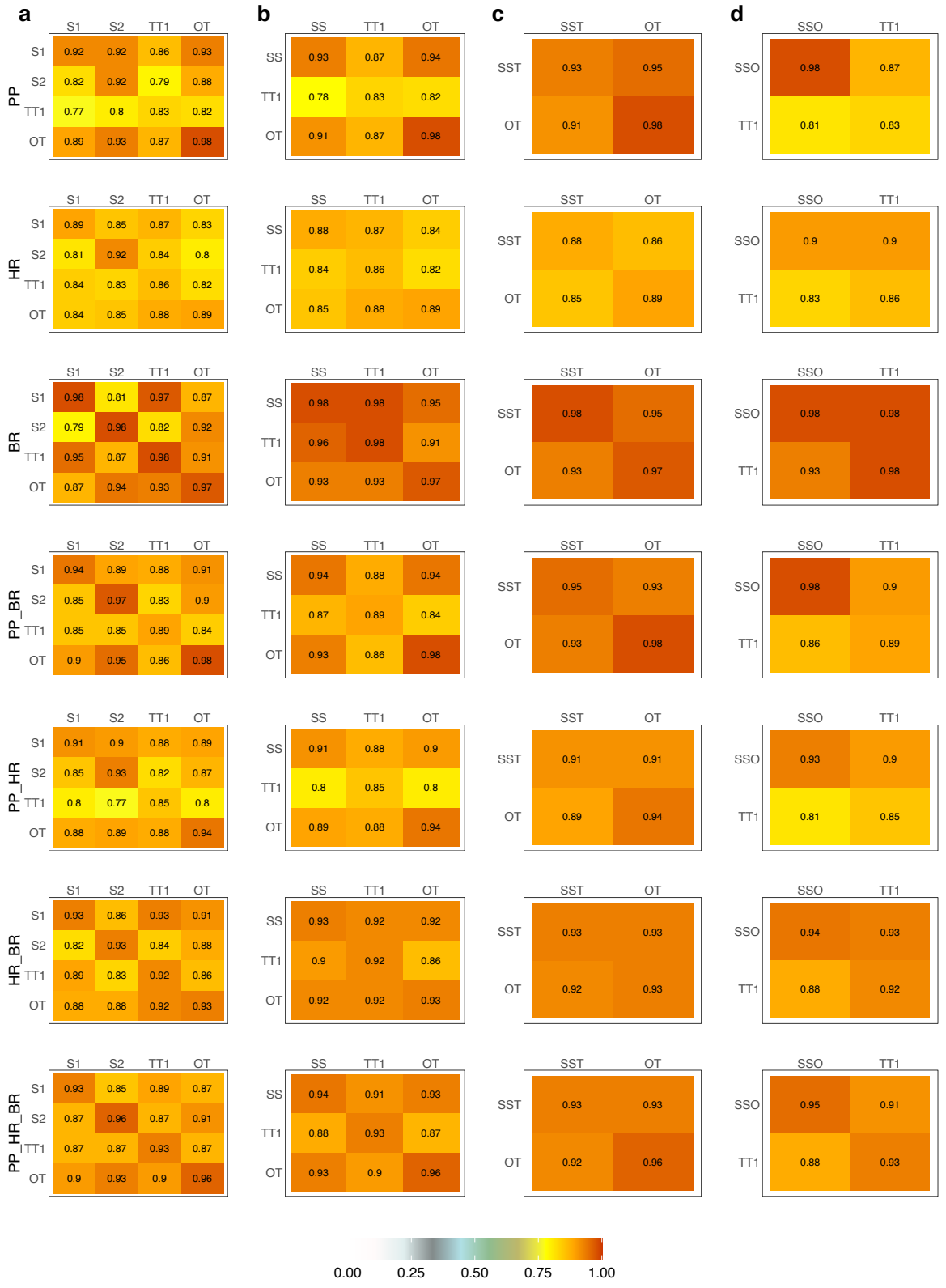

Fig. 5: **Time dependent classification results in the form of combinatorial study matrices.** Matrix rows indicate training studies while matrix columns indicate testing studies. The symbols in the diagonal cells denote the studies that partake in the training-testing combinations. The numbers in the off-diagonal cells are the cross-study AUCs. Panel rows show classification results associated with sensing modalities PP, HR, BR and their combinations thereof. Panel columns show classification results associated with different study inclusion methods in the training sets. **Panels a.** Training sets include single studies. **Panels b.** Training sets include multiple driving controlled studies. **Panels c.** Training sets include driving studies across the realism axis. **Panels d.** Training sets include studies across the domain axis.

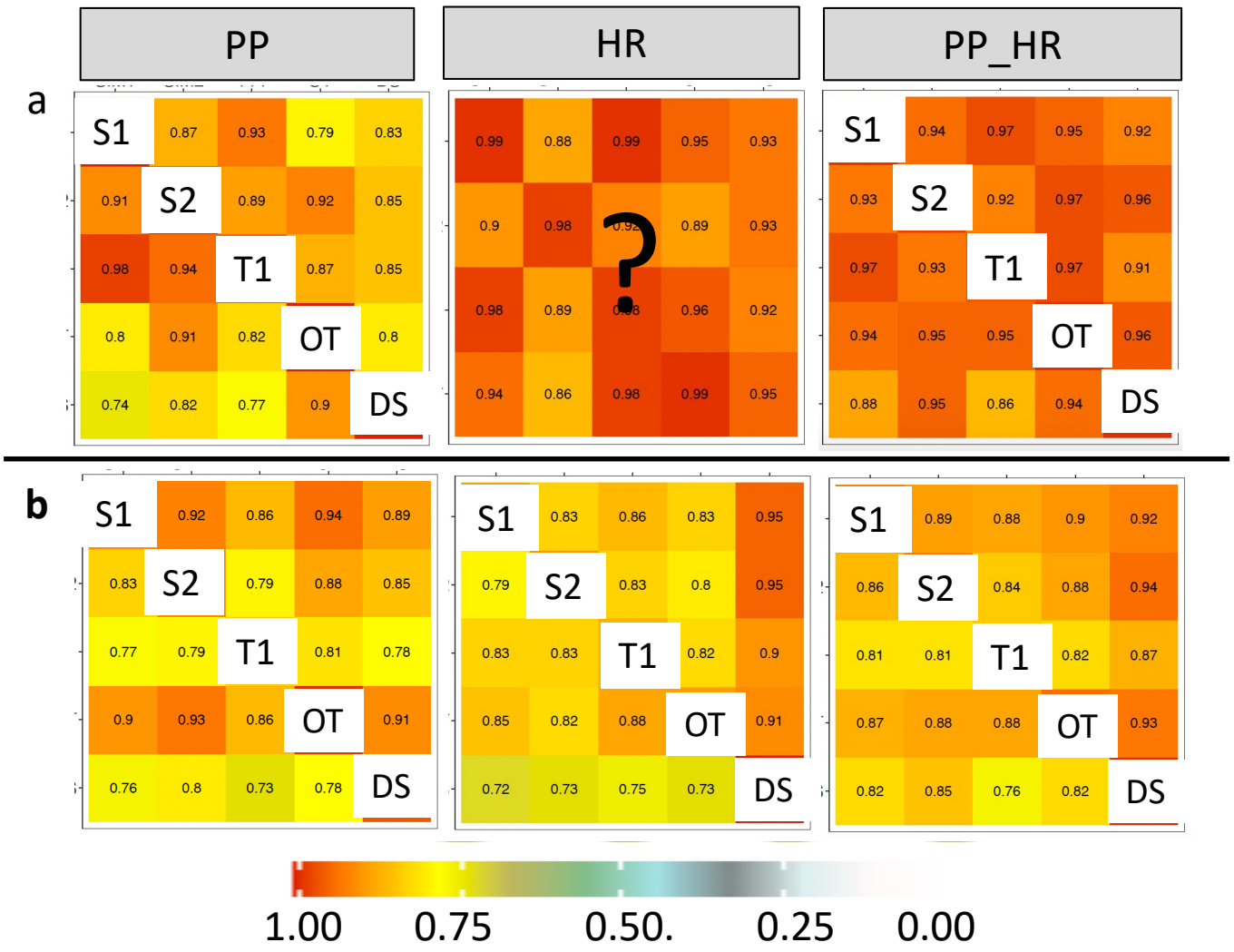

Fig. 6: **Complete matrix. Panels a.** Time independent. **Panels b.** Time dependent.