Preliminary evidence of physiological markers of core affective experience: Results from a big team machine learning competition

Nicholas A. Coles1†, Bartosz Perz2, Maciej Behnke3, Johannes Eichstaedt1, Ngoc Tu Vu4, Soo-Hyung Kim4, Van Thong Huynh5, Guangyi Zhang6, Arpit Upadhyay7, Shubham Pandey8, Rushi Chavda9, Nicolás M. Bruno10, Federico Zamberlan10, Tomás A. D'Amelio10, Leandro A. Bugnon10, Enzo R. Tagliazucchi11,12, Felix Dollack13, Monica Perusquia-Hernandez13, Hideaki Uchiyama13, Huakun Liu13, Xin Wei13, Kiyoshi Kiyokawa13, Alessia Iancarelli14, Jamie C. Chiu15, Claire Whiting15, Henna I. Vartiainen15, Dan-Mircea Mirea15, SHARANYAK PODDER16, Yanling Li17, Zachary F. Fisher17, Young Won Cho17, Luis Roberto Mercado Diaz18, Hugo F. Posada-Quintero18, Saurabh Hinduja19, Maneesh Bilalpur19, Stanislaw Saganowski2  
  
1Stanford University, USA, 2Wroclaw University of Science and Technology, Poland, 3Adam Mickiewicz University, Poland, 4Chonnam National University, South Korea, 5FPT University, Vietnam, 6Harvard Medical School, USA, 7IIT Bombay, India, 8Indian Institute of Technology Bombay Mumbai, India, 9Indian Institute of Technology, Bombay, India, 10INFINA – UBA – CONICET , Argentina, 11INFINA – UBA – CONICET, Argentina, 12BrainLat – UAI, Argentina, 13Nara Institute of Science and Technology, Japan, 14Northeastern University, USA, 15Princeton University, USA, 16Queen Mary University of London, United Kingdom, 17The Pennsylvania State University, USA, 18University of Connecticut, USA, 19University of Pittsburgh, USA  
  
†Correspondence should be addressed to Nicholas A. Coles; E-mail: ncoles@stanford.edu

**Dear collaborators: Please check!** People who seem potentially eligible for authorship but never submitted authorship information.

|  |  |  |  |
| --- | --- | --- | --- |
| Arpit Upadhyay | Jorge Ivan Padilla | Xiaoyue Xiong | Ziqing Yang |
| Liza Jivnani | Julian Tejada | Yuqi Shen | Erik Nook |
| Shaun Canavan | Raydonal Ospina | Tanming Cui | Isabel Berwian |
| Fernando Marmolejo-Ramos | Chirag Raman | Kieran McVeigh | Ali Etemad |
| Javier Pinzón-Arenas | Linying Ji | Yiyu Wang | Kleanthis Avramidis |
| Carlos Barrera-Causil | Sy-Miin Chow | Houwei Cao | Vu Tu Huynh |
| Soo-Hyung |  |  |  |

# Abstract

Many theorists posit that subjective emotional experiences have a physiological basis – often suggesting that these experiences are built off a potentially complex set of afferent feedback from the peripheral nervous system (PNS). The potential complexity of this relationship, along with outstanding theoretical disagreements about its nature, have prompted researchers to use state-of-the-art methods in machine learning. Unfortunately, issues with comparability and transparency limit what we can learn from past efforts. To address this barrier, we organized a machine learning competition, wherein 12 teams (50 researchers) competed to predict self-reported core affective experiences using a multi-modal set of PNS features. In 100% of tests, teams outperformed a baseline model that made random predictions. In 46% of tests, teams also outperformed a baseline model that based predictions on averaged ratings from training datasets, albeit by a usually small magnitude. In a follow-up experiment, three models subjectively judged most promising by independent evaluators exhibited lower prediction accuracy when re-tested with simulated physiological randomness. These results provide preliminary evidence of physiological markers of core affective experiences that can be captured and studied via machine learning. However, results also uncovered a methodological challenge: multiplicative constraints on generalizability. Inferences about the accuracy and theoretical implications of machine learning models depended not only on their architecture, but also how they were trained, tested, and evaluated. For example, some modeling approaches achieved higher performance when tested on observations from different (vs. the same) subjects seen during training. Such results could be interpreted as evidence of biologically innate links that are sensitive to context. However, such conclusions would be premature because other modeling approaches exhibited the opposite pattern. Nonetheless, by standardizing these methodological decisions and crowdsourcing the development of models, our project illustrates how big team science can be leveraged to understand and navigate these constraints on generalizability.

*Keywords:* emotion AI, physiology, machine learning, generalizability, big team science, affective computing

# Introduction

All throughout their lives, people experience phenomenological states they call “emotions”1. Many theorists have posited that these emotional experiences have a *physiological basis* – often suggesting that they are [partially or fully] built off afferent feedback from the peripheral nervous system (PNS)2–12. If true, such accounts provide clues not only about *how* emotional experiences arise, but also *why* such a capacity evolved. For example, imagine that PNS activity both (a) responds to homeostatic and/or allostatic demand, and (b) informs the conscious experience of emotion. If true, emotional experiences may have evolved to allow people to monitor and regulate these physiological states3,13.

If PNS activity is the physiological basis of emotional experience, it should be predictive of peoples’ subjective emotion reports. This idea is challenging to evaluate because the relationship between PNS activity and emotion experiences may be ultra-complex, involving a large number of afferent signals14, non-linear and time-lagged relationships15, and high-order interactions. Furthermore, although theorists may agree that emotional experience is partially built off PNS activity, they do not necessarily agree on the specification of such relationship. For example, there are outstanding questions about the extent to which emotion-related PNS activity is similar across people, emotion types, contexts, and time16. However, advancements in machine learning raise an interesting possibility: in the absence of well-specified theories, is it possible to model a potentially ultra-complex physiological basis of emotional experience via machine learning?

Evaluating the extent to which past machine learning efforts have achieved this goal is challenging because of issues with *commensurability* (i.e., comparability) and transparency. More specifically, it is challenging to compare results across studies because researchers often use (a) different datasets (e.g., different populations, sets of sensors, and data processing procedures), (b) different outcomes (e.g., observer ratings vs. self-reported emotion; a focus on core affect vs. discrete emotions), (c) different performance benchmarks (e.g., different tasks and criteria for evaluating accuracy), and (d) different model validation procedures17,18.

As an illustrative example, consider the challenges of comparing two influential studies by Picard et al.19 and Haag et al.20 (Table 1). On one hand, Picard et al. (a) used PNS data from a single participant who completed an emotion self-elicitation task (e.g., guided imagery) to (b) predict the emotion targeted by the task. They (c) defined accuracy as the percentage of times the model correctly predicted which emotion was targeted, and (d) tested their model using leave-one-out cross-validation, concluding 81% classification accuracy. On the other hand, Haag et al. (a) used PNS data from a single participant who viewed emotional photos, to (b) predict how other participants rated those photos. They (c) defined accuracy as the percentage of times the predicted value fell within a range of ratings from other participants, and (d) tested their model using hold-out validation, concluding 90% and 97% accuracy for valence and arousal reports respectively. As was common in the early 2000’s, it does not appear that materials and code from those projects were made openly available, making it difficult to carefully inspect the models, compare their results, and test the generalizability of their performance21.

**Table 1.** The results of past machine learning efforts (two shown here) are difficult to compare because they typically used different datasets, outcomes, benchmarks, and testing procedures. The present work standardizes and/or systematically varies those methodological details.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Citation** | **Dataset** | **Outcome** | **Benchmark** | **Testing Procedure** |
| Picard et al. (2001) | N = 1 subject completes emotion self-elicitation task | Emotion targeted by the self-elicitation task | Percentage of times the model correctly predicted which emotion was targeted | Leave-one-out cross-validation |
| Haag et al. (2004) | N = 1 subject views emotional photos | External ratings of emotional photos | Percentage of times predicted value fell within a range of the external ratings | Hold out validation |
| Present work with 12 teams | N = 30 subjects view emotional videos | Self-reported core affect | Absolute prediction error (compared to two baseline models) | Across-time, across-subject, across-emotion, and across-induction validation |

The purpose of the present work is to evaluate – in a manner that promotes commensurability and transparency – the extent to which machine learning methods can model theorized links between PNS activity and emotion self-reports. To do so, we organized a big team science competition involving 12 teams with a total of 50 researchers. For the competition, we used an openly available dataset, wherein 30 subjects encountered two video inductions of fear, boredom, relaxation, and amusement (8 inductions total)22. Throughout the inductions, participants used a joystick to provide continuous ratings of valence (negativity vs. positivity) and arousal (calm vs. excited). Simultaneously, a multi-modal set of PNS measures was collected (e.g., electrocardiography, electrodermal activity, and respiration). Notably, both the core affect reports and PNS signals were collected moment-to-moment (20 and 1000 samples per second, respectively). This provided teams with a large set of observations with high temporal resolution. Using these observations, teams were challenged to leverage PNS features to predict core affect reports via machine learning. See SI for summaries of each team’s approach.

To illustrate the practical and theoretical implications of decisions about model testing, we used four different approaches (see Materials and Methods for more information). For *across-subject validation*, models were (a) trained on one set of participants and (b) tested on a different set of participants. For *across-emotion validation*, models were (a) trained on inductions of three affective states, and (b) tested on inductions of a fourth affective state. For *across-induction validation*, models were (a) trained on one induction of each affective state, and (b) tested on a second induction of those same affective states. For *across-time validation*, models were (a) trained on the beginning of the videos and (b) tested on later parts of the videos. Notably, these validation approaches roughly map onto theoretical debates about the extent to which links between PNS activity and emotional experience are biologically innate (across-subject validation), emotion specific (across-emotion validation), similar across contexts (across-induction validation), and stable throughout the course of an emotional event (across-time validation)16. As described in the Results section, two competition organizers reviewed teams’ performance and underlying code to identify three approaches that seemed particularly promising. Their models were then re-tested with simulated physiological randomness to further evaluate the role of PNS activity.

The outcome of interest was the absolute value of the prediction error for self-reports of the two dimensions of core affect: valence and arousal. We chose to focus on self-reports because it most closely maps onto our construct of interest: peoples’ subjective emotional experiences. To illustrate the impact of decisions about benchmarking, we compared teams to two baseline models: (1) a *random baseline model* that made random (within the range of the measure) predictions about core affect, and (2) a *mean baseline model* that uniformly predicted core affect ratings as the mean of the ratings observed in the training dataset. At the end of the competition, all data, materials, and code were uploaded to a series of openly available repositories.

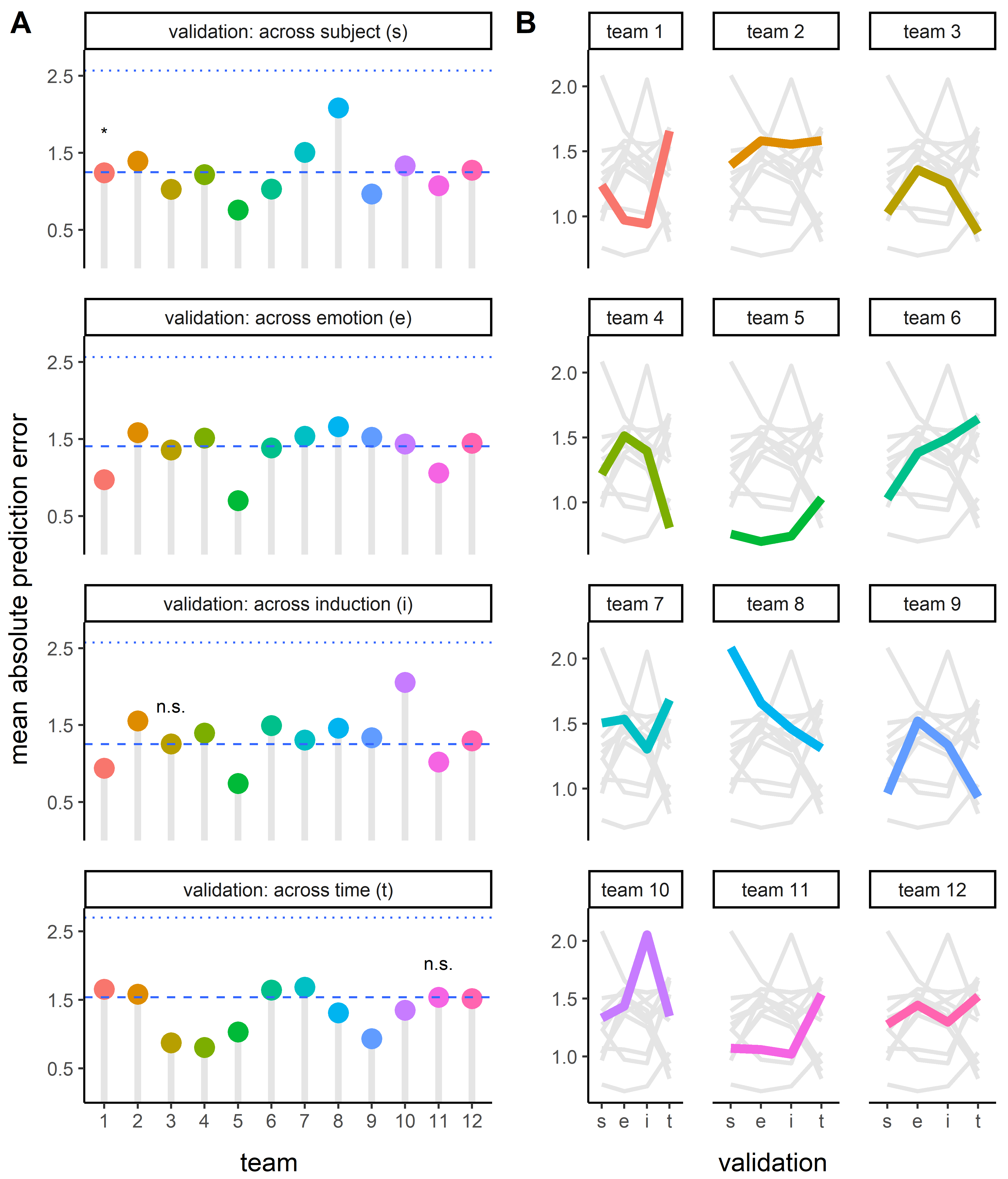
# Results

To examine each team’s accuracy, we used mixed-effect regression. Mixed-effect regression is a robust analytic approach that can be used to accommodate non-independent observations, which often lead to inflated error rates in traditional linear regression or data loss when averaged over 23–25. For each team, we regressed absolute prediction error as a function of (a) whether the prediction came from a model developed by the team, the random baseline model, or the mean baseline model, (b) the validation approach used for testing, (c) a higher-order interaction between model source and validation approach, and (d) random intercepts for each subject and video in the dataset. Random intercepts were included to accommodate non-independent observations from the same participants and videos. For each validation approach, we used model-derived pairwise contrasts to estimate and test the significance of the mean difference (*MD*) in the absolute prediction error between the model developed by the team and each baseline model (Figure 1A).

In every validation approach, all 12 teams made more accurate predictions than a *random baseline model* (1.89 < *MD* > 0.48, all *z* > 130.31, all *p* < .001). However, teams did not always make more accurate predictions than a *mean baseline model*. Seven teams (58%) outperformed the mean baseline model in across-subject (0.49 < *MD* > 0.01, all z > 2.44, all *p* < .05) and across-time validation (0.73 < *MD* > 0.02, all *z* > 5.13, all *p* < .001); five teams (42%) outperformed the mean baseline model in across-emotion validation (0.70 < *MD* > 0.02, all *z* > 5.81, all *p* < .001); Three (25%) teams outperformed the mean baseline model in across-induction validation (0.51 < *MD* > 0.23, all *z* > 62.58, all *p* < .001).

As further illustrated in Figure 1B, the manner in which prediction accuracy varied across validation approaches differed across teams. For example, team 1 achieved lower prediction error in across-subject vs. across-time validation. Team 4 exhibited the opposite pattern. These results highlight how constraints on generalizability are *multiplicative*. Inferences about the ability to leverage machine learning to capture links between PNS activity and emotion reports depends not only on (a) what architecture is used, but also (b) how the models are tested.

**Figure 1.** *Panel A:*Absolute error of core affect predictions (y-axis) made by models developed by 12 teams (x-axis). Validation approaches (panels) are visualized separately. Models are compared to a random baseline (lower dotted line) and mean baseline (higher dashed line). *Panel B*: Re-illustration of team-specific (panel) error (y-axis) across validation approaches (x-axis).



*Panel A Note.* All *p* < .001 unless otherwise indicated; *\** denotes *p* < .05; n.s.denotes *p* > .05

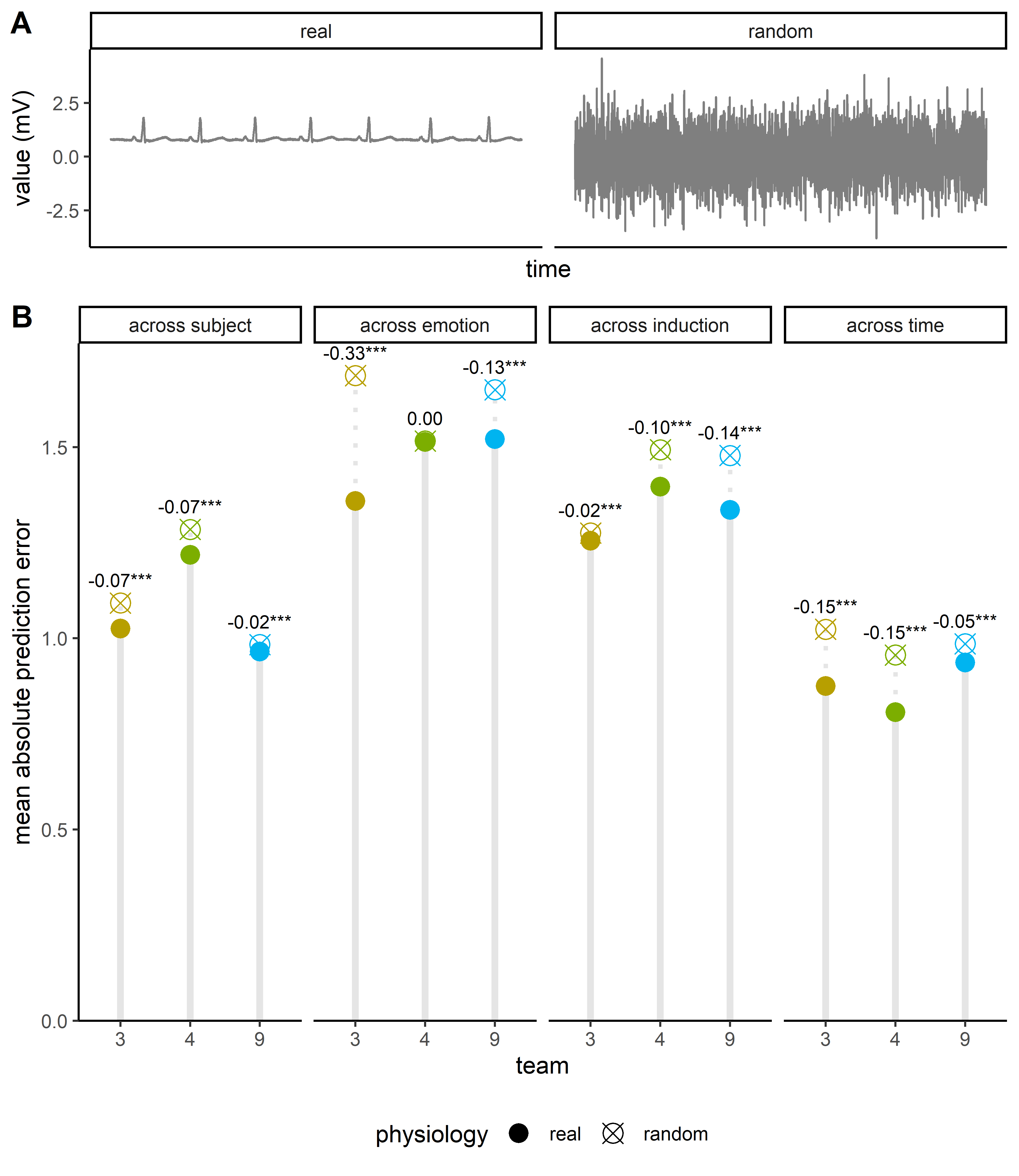
## Further evidence of the role of PNS activity

The above results provide preliminary evidence that machine learning models uncovered links between PNS activity and core affect reports. Their predictions always outperformed random guessing, but they did not always outperform a mean baseline model that uniformly predicted core affect ratings as the mean of the ratings observed in the training dataset (Figure 1A). This raises questions about the extent to which their predictions were driven by (a) theory-relevant potential links between PNS activity and core affect reports vs. (b) the recovery of theory-irrelevant averages of emotion reports in the training dataset.

To further investigate the role of PNS activity, competition organizers reviewed the openly available code for top-performing submissions and [subjectively] chose three that seemed particularly promising. The teams’ models were then re-tested on the same data with one change: measures of PNS activity were replaced with simulated physiological randomness (Figure 2A).

Using mixed-effect regression, we regressed each team’s absolute prediction error as a function of (a) the validation approach used for testing, (b) whether the test data contained real or simulated physiological randomness, (c) a higher-order interaction between validation approach and whether the physiology input was real, and (d) random intercepts for each subject and video in the dataset. In all but one case (*MD* = 0.00, *z* = -1.01, *p* = .31), the teams’ prediction accuracy decreased when tested on simulated physiological randomness, -0.33 < *MD* > -0.02, all *z* < -6.56, all *p* < .001 (Figure 3). In other words, accuracy decreased in 93% of tests with simulated physiological randomness, providing further evidence that at least some models considered PNS signals to predict self-reported core affect.

**Figure 2.** *Panel A:*Example of real electrocardiography signal vs. simulated physiological randomness. *Panel B.* Absolute error of core affect predictions (x-axis) made by models developed by 3 teams (y-axis) when tested on real physiology (circle) vs. simulated physiological randomness (crossed circle). Results are visualized separately for four validation approaches (panels). Model-derived mean differences are reported.



*Panel B Note.* \*\*\*denotes *p* < .001.

# Discussion

Our results provide evidence that machine learning can be leveraged to model potentially complex physiological markers of core affective experiences. In four different tests, models developed by 12 teams of researchers uniformly achieved accuracy that exceeded what would be expected by mere random guessing. About half the time, the accuracy of these models also exceeded what would be expected by merely leveraging mean ratings from training datasets. Further evidence of the role of PNS activity comes from follow-up tests with simulated physiological randomness, which nearly uniformly caused models from three selected teams to become less accurate. However, the magnitude of differences in prediction accuracy in many tests was small, illustrating extensive opportunity for future improvements.

Our results are consistent with theories that posit a physiological basis of emotional experience2–12. However, such results are also consistent with competing theories that posit that these links are spurious – e.g., that PNS activity and emotional experience are independent byproducts of an upstream neural process (e.g., in the thalamic region)26,27. Although there have been advancements in machine learning methods for evaluating causal relationships28, the present work was not designed to provide strong tests. For example, team 4’s model (described elsewhere29) indicated that core affect reports were best predicted by PNS activity that preceded (vs. co-occurred or succeeded) the emotion reports. Establishing the temporal precedence of PNS activity would bolster claims causality25. However, such patterns may be alternatively driven by delays in how quickly participants can report their core affective states. To address such limitations, future machine learning researchers may wish to consider evidence from experiments with human subjects. For example, causal claims are bolstered by studies documenting the effects of experimentally manipulated PNS activity on participants’ emotion reports30–35.

More practically, our results underscore challenges that past and future researchers face with commensurability and generalizability in research that seeks to leverage machine learning to predict and understand emotion36,37. Our results indicate that differences in how models are developed, benchmarked, *and* tested can impact researchers’ conclusions. For instance, focusing on a random baseline (vs. mean baseline) leads to a more optimistic interpretation of models’ accuracy – as does focusing on across-time validation (vs., for example, across-emotion validation). These results have implications for ongoing discussions about the potential benefits (e.g., unobtrusive measurement of internal emotional states) and harms (e.g., inaccurate predictions) of emotion recognition technologies38–42 – such as those being pursued by an emotion AI industry recently valued at $20+ billion43.

Even more challenging is our observation that the aforementioned constraints on generalizability are *multiplicative* (i.e., interactive)37. For example, the accuracy of predictions depended both on the modeling approach and the validation approach. However, it is inadvisable to make broad conclusions about the impact of any one of these decisions because they had an *interactive* effect on prediction accuracy. For example, team 1 achieved lower prediction error in across-subject vs. across-time validation. Some researchers may be tempted to conclude that this provides evidence that links between PNS activity and emotion reports vary more within-persons than between-persons. This conclusion, if true, would bolster claims that such links are biologically-innate44,45 but sensitive to context. The conclusion would also suggest that future data collection and modeling efforts should focus less on the diversity of the sampled population and more on the diversity of the emotional contexts they encounter. However, such conclusions are premature when considering that differences in both model *and* validation approaches have interactive effects on prediction accuracy. Indeed, models developed by other teams (e.g., team 8) exhibited the opposite pattern: that prediction error is lower in across-time vs. across-subject validation. Such results could be interpreted as evidence *against* claims of biological innateness and would underscore the importance of collecting diverse participant samples46. These multiplicative constraints of generalizability will be important to keep in mind as researchers increasingly leverage machine learning not only to predict emotion – but to try to evaluate theoretical claims about the nature of the modeled process47–49.

Although our work highlights challenges with commensurability and multiplicative constraints on generalizability, it also provides proof-of-concept for a potential methodological response: big team science50–53. Big team science effectively allowed us to leverage the wisdom-of-crowds to test a challenging theoretical question in emotion research. Standardizing methodological decisions about data sources, benchmarks, and testing procedures permitted cleaner comparisons of teams’ different approaches. Further introducing *systematic variation* in specific methodological decisions (e.g., testing procedures) allowed us to empirically examine the extent to which these decisions constrain the generalizability of inferences. Last, requiring that teams make their materials and code openly available allowed us (and future researchers) to further inspect teams’ models, reproduce their solutions, and identify approaches that seem most promising for follow-up research21.

Our examination of the utility of big team science in emotion recognition research also yielded lessons about ways that future collaborative efforts can be improved. For instance, feasibility constraints prohibited the competition organizers from closely examining the code for all submissions and working closely with teams to investigate any concerns raised during code review. Developing protocols and best practices for peer code review would have better enabled the crowdsourcing of this task54–57. Teams that entered the competition also faced resource constraints, with many expressing that they would have benefited from more time to work on the challenge, bigger datasets, and/or access to more powerful computing resources. Recent pushes to collaborate on dataset development50–53 and provide access to shared computing resources58 may prove instrumental in helping researchers overcome those barriers. Despite existing constraints, our work raises an exciting possibility: advancements in machine learning and collaborative research methods provide researchers with new tools for tackling ultra-complex questions in affective science and beyond.

# Materials and Methods

For the competition, we used the Continuously Annotated Signals of Emotion dataset (Figure 1B)22. In this dataset, 30 subjects encountered two video inductions of fear, boredom, relaxation, and amusement. In total, there were 8 inductions, each 2-3 minutes in length. Throughout the inductions, participants used a joystick to navigate a continuous, 9-point, two-dimensional grid measuring subjective core affective experience: valence (negativity vs. positivity) on the x-axis and arousal (calm vs. excited) on the y-axis. Simultaneously, a multi-modal set of PNS measures were collected: electrocardiography, blood volume pulse, electrodermal activity, respiration, skin temperature, and electromyography activity over the zygomaticus major, corrugator supercilii, and trapezius muscles.

For the competition, we sought to recruit up to 15 teams based on the availability of funds. Teams were eligible to participate in the challenge if they agreed to (a) make their code openly available, (b) collaborate on a manuscript describing the challenge results, and (c) not cheat (e.g., look for the original dataset). Eighteen teams completed the application to join the challenge, which asked them to report previous experience with machine learning challenges, the number of papers published in this domain, their planned approach to the challenge, and the CV of members of their teams. The competition organizers selected teams based on their averaged evaluations of their (a) experience with machine learning challenges (N.C.), (b) expertise in emotion research (M.B.), and/or (c) planned approach to the challenge (S.S.). These evaluations were made independently and then discussed as a group. Of the 15 teams that were invited to compete, 3 dropped out due to difficulties they encountered during the challenge. In total, 12 teams completed the challenge using a variety of modeling approaches (see SI Table 1)29,59,60.

Teams received $300 for completing the competition, and those who developed models judged to be most promising received a $200 bonus. This payment structure was designed to incentivize balance incentives for effort ($300 for all teams) and performance ($200 for top performing teams). To subjectively identify the most promising submissions, two competition organizers (S.S., B.P.) worked together to examine (a) how well the teams performed by inspecting their root mean square error, and (b) (if top performing) their underlying code. This two-step process allowed them to consider both performance and perceived methodological soundness when choosing which models to further evaluate.

For model training and testing, we created four validation scenarios that roughly map onto theoretical debates about the extent to which links between PNS activity and emotional experience are biologically innate (across-subject validation), emotion specific (across-emotion validation), similar across contexts (across-induction validation), and stable throughout the course of an emotional event (across-time validation)

For across-subject validation, we used leave-N-subjects-out validation. Participants were randomly divided into five folds. Teams trained models on four folds and tested models on a fifth fold. This was repeated for each combination of folds.

For across-emotion validation, we used a leave-one-emotion-out validation approach. As a reminder, four emotions were targeted (via videos) in the original dataset: amusement, fear, boredom, and relaxation. We created one fold for each targeted emotion (four folds total). Teams trained models on data from three targeted emotions and tested models on data from a fourth targeted emotion. This was repeated for each combination of folds.

For across-induction validation, we used a hold-out validation approach. As a reminder, each targeted emotion was induced through two different videos. For each emotion, teams trained models on data from one video and tested models on data from the second video.

For across-time validation, we used a hold-out validation approach focused on chronology. For each participant, we divided the data from each emotion induction into training and test sets based on time. Teams trained models on the beginning of the inductions and tested the models on the later parts of the inductions.

Teams were permitted to use different modeling approaches. Upon request, teams were also permitted to test their models up to four times: three optional tests on a subset (50%) of test data and a fourth test on all test data. In follow up tests probing the role of PNS features (Figure 1D), the PNS data in the testing files were replaced with simulated physiological randomness: *N*(μ = 0, σ = 1).

For benchmarking, we focused on two relatively simple baseline models: (1) a *random baseline model* that made random (within the range of the measure) predictions about core affect, and (2) a *mean baseline model* that uniformly predicted core affect ratings as the mean of the ratings observed in the training dataset. However, other baseline models could certainly be considered, such as ones that calculate tailored subject-specific mean ratings.

**Figure 3.** Overview of the validation structure.

# References

1. Damasio, A. R. *Descartes’ error: Emotion, reason, and the human brain*. (Harper Perennial, 1994).

2. Cacioppo, J. T., Berntson, G. G. & Klein, D. J. What is an emotion? The role of somatovisceral afference, with special emphasis on somatovisceral ‘illusions’. in *Review of personality and social psychology: Emotion and social behavior* (ed. Clark, M. S.) 63–98 (Sage Publications, Inc., 1992).

3. Damasio, A. & Carvalho, G. B. The nature of feelings: Evolutionary and neurobiological origins. *Nat. Rev. Neurosci.* **14**, 143–152 (2013).

4. Tomkins, S. *Affect Imagery Consciousness: Vol 1 The Positive Affects*. (Springer, 1962).

5. Wood, A., Rychlowska, M., Korb, S. & Niedenthal, P. Fashioning the face: Sensorimotor simulation contributes to facial expression recognition. *Trends Cogn. Sci.* **20**, 227–240 (2016).

6. James, W. What is an emotion? *Mind* **9**, 188–205 (1884).

7. James, W. Discussion: The physical basis of emotion. *Psychol. Rev.* **1**, 516–529 (1894).

8. Laird, J. D. & Bresler, C. The process of emotional experience: A self-perception theory. in *Review of Personality and Social Psychology: Emotion* (ed. Clark, M. S.) 213–234 (1992).

9. Lange, C. G. *Om sindsbevaegelser; et psyko-fysiologisk studie*. (Lund, 1885).

10. Levenson, R. W., Ekman, P. & Friesen, W. V. Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology* **27**, 363–384 (1990).

11. Russell, J. A. A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**, 1161–1178 (1980).

12. Scherer, K. R. & Moors, A. The emotion process: Event appraisal and component differentiation. *Annu. Rev. Psychol.* **70**, 719–745 (2019).

13. Barrett, L. F. *How emotions are made: The secret life of the brain*. (Pan Macmillan, 2017).

14. Birch, R. The peripheral nervous system: anatomy and function. in *Peripheral Nerve Injuries: A Clinical Guide* 1–67 (Springer, 2013).

15. Asutay, E. *et al.* Affective calculus : The construction of affect through information integration Affective calculus : The construction of affect through information integration over time. *Emotion* (2019) doi:10.1037/emo0000681.

16. Lindquist, K. A., Siegel, E. H., Quigley, K. S. & Barrett, L. F. The hundred-year emotion war: Are emotions natural kinds or psychological constructions? Comment on Lench, Flores, and Bench (2011). *Psychol. Bull.* **139**, 255–263 (2013).

17. Calvo, R. A. & D’Mello, S. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Trans. Affect. Comput.* **1**, 18–37 (2010).

18. Saganowski, S., Perz, B., Polak, A. & Kazienko, P. Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review. *IEEE Trans. Affect. Comput.* **12**, 1–1 (2022).

19. Picard, R. W., Vyzas, E. & Healey, J. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**, 1175–1191 (2001).

20. Haag, A., Goronzy, S., Schaich, P. & Williams, J. Emotion recognition using bio-sensors: First steps towards an automatic system. in *Tutorial and research workshop on affective dialogue systems* 36–48 (Springer, 2004).

21. Haibe-Kains, B. *et al.* Transparency and reproducibility in artificial intelligence. *Nature* **586**, E14–E16 (2020).

22. Sharma, K., Castellini, C., van den Broek, E. L., Albu-Schaeffer, A. & Schwenker, F. A dataset of continuous affect annotations and physiological signals for emotion analysis. *Sci. Data* **6**, 1–13 (2019).

23. Brauer, M. & Curtin, J. J. Linear mixed-effects models and the analysis of nonindependent data : A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychol. Methods* **23**, 389–411 (2018).

24. Schielzeth, H. *et al.* Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods Ecol. Evol.* **11**, 1141–1152 (2020).

25. Frank, M. C. *et al.* *Experimentology: An open science approach to experimental psychology methods*. (Boston, MA: MIT Press, 2023).

26. Cannon, W. B. Again the James-Lange and the thalamic theories of emotion. *Psychol. Rev.* **38**, 281–295 (1931).

27. Cannon, W. The James-Lange theory of emotions: A critical examination and an alternative theory. *Am. J. Psychol.* **39**, 106–124 (1927).

28. Lagemann, K., Lagemann, C., Taschler, B. & Mukherjee, S. Deep learning of causal structures in high dimensions under data limitations. *Nat. Mach. Intell.* **5**, 1306–1316 (2023).

29. D’Amelio, T. A., Bruno, N. M., Bugnon, L. A., Zamberlan, F. & Tagliazucchi, E. Affective computing as a tool for understanding emotion dynamics from physiology: A predictive modeling study of arousal and valence. in *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* 1–7 (IEEE, 2023).

30. Coles, N. A. *et al.* A multi-lab test of the facial feedback hypothesis by the Many Smiles Collaboration. *Nat. Hum. Behav.* **6**, 1731–1742 (2022).

31. Coles, N. A. *et al.* Fact or artifact? Demand characteristics and participants’ beliefs can moderate, but do not fully account for, the effects of facial feedback on emotional experience. *J. Pers. Soc. Psychol.* **124**, 287–310 (2022).

32. Coles, N. A., Larsen, J. T. & Lench, H. C. A meta-analysis of the facial feedback literature: Effects of facial feedback on emotional experience are small and variable. *Psychol. Bull.* **145**, 610–651 (2019).

33. Körner, R., Röseler, L., Schütz, A. & Bushman, B. J. Dominance and prestige: Meta-analytic review of experimentally induced body position effects on behavioral, self-report, and physiological dependent variables. *Psychol. Bull.* **148**, 67–85 (2022).

34. MacCormack, J. K. & Lindquist, K. A. Bodily contributions to emotion: Schachter’s legacy for a psychological constructionist view on emotion. *Emotion Review* vol. 9 36–45 (2017).

35. MacCormack, J. K. & Lindquist, K. A. Feeling hangry? When hunger is conceptualized as emotion. *Emotion* **19**, 301–319 (2019).

36. Yarkoni, T. The generalizability crisis. *Behav. Brain Sci.* (2020).

37. Almaatouq, A. *et al.* Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behav. Brain Sci.* 1–55 (2022).

38. Hernandez, J. *et al.* Guidelines for assessing and minimizing risks of emotion recognition applications. *2021 9th Int. Conf. Affect. Comput. Intell. Interact.* (2021).

39. Ong, D. C. An ethical framework for guiding the development of affectively-aware artificial intelligence. *2021 9th Int. Conf. Affect. Comput. Intell. Interact.* (2021).

40. Bryant, D. A. & Howard, A. Age bias in emotion detection: An analysis of facial emotion recognition performance on young, middle-aged, and older adults. *Proc. 2021 AAAI/ACM Conf. AI, Ethics, Soc.* 638–644 (2021).

41. Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M. & Pollak, S. D. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychol. Sci. Public Interes.* **20**, 1–68 (2019).

42. Boyd, K. L. & Andalibi, N. Automated emotion recognition in the workplace : How proposed technologies reveal potential futures of work. *PACM Human-Computer Interact.* **5**, 95:1-95:37 (2023).

43. Telford, T. ‘Emotion detection’ AI is a $20 billion industry. New research says it can’t do what it claims. *Washington Post* (2019).

44. Ekman, P. Basic Emotions. in *Handbook of Cognition and Emotion* (eds. Dlgleish, T. & Power, M.) 45–60 (John Wiley & Sons Ltd., 1999).

45. Tracy, J. L. & Randles, D. Four models of basic emotions: A review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt. *Emot. Rev.* **3**, 397–405 (2011).

46. Hussein, A. *et al.* Ethical AI in facial expression analysis: Racial bias. *Signal, Image Video Process.* **17**, 399–406 (2023).

47. Hoemann, K. *et al.* Context‑aware experience sampling reveals the scale of variation in affective experience. *Sci. Rep.* **10**, 1–16 (2020).

48. Barrett, L. F. Debate about universal facial expressions goes big. *Nature* **589**, 200–201 (2021).

49. Azari, B. *et al.* Comparing supervised and unsupervised approaches to emotion categorization in the human brain, body, and subjective experience. *Sci. Rep.* **10**, 1–17 (2020).

50. Coles, N. A., Hamlin, J. K., Sullivan, L. L., Parker, T. H. & Altschul, D. Build up big-team science. *Nature* **601**, 505–507 (2022).

51. Forscher, P. S. *et al.* The benefits, barriers, and risks of big-team science. *Perspect. Psychol. Sci.* **18**, 607–623 (2023).

52. Coles, N. A., Debruine, L. M., Azevedo, F., Baumgartner, H. A. & Frank, M. C. ‘Big team’ science challenges us to reconsider authorship. *Nat. Hum. Behav.* **7**, 665–667 (2023).

53. Baumgartner, H. *et al.* How to build up big team science: A practical guide for large-scale collaborations. (2023).

54. Obels, P., Lakens, D., Coles, N. A. & Gottfried, J. Analysis of open data and computational reproducibility in registered reports in psychology.

55. Hardwicke, T. E. & Vazire, S. Transparency Is Now the Default at Psychological Science. *Psychological Science* 09567976231221573 (2023).

56. Cracking the code review process. *Nat. Comput. Sci.* **2**, 277 (2022).

57. Ivimey‐Cook, E. R. *et al.* Implementing code review in the scientific workflow: Insights from ecology and evolutionary biology. *J. Evol. Biol.* **36**, 1347–1356 (2023).

58. Lohr, S. Universities and Tech Giants Back National Cloud Computing Project. *New York Times* (2020).

59. Pinzon-Arenas, J. O. *et al.* Deep learning analysis of electrophysiological series for continuous emotional state detection. in *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* 1–8 (IEEE, 2023).

60. Dollack, F. *et al.* Ensemble learning to assess dynamics of affective experience ratings and physiological change. in *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* 1–8 (IEEE, 2023).

**Supplementary Table 1.** Descriptions of each team’s name and modeling approach.

|  |  |  |
| --- | --- | --- |
| **Team** | **Team name** | **Description of approach** |
| 1 | AffectiveBulls | Physiological signals were used to derive a univariate time-series representation of all signals through a weighted sum of all signals, with normalized variance used as weights. This signal was split into windows of fixed duration and used for training person-specific neural networks that predicted valence and arousal ratings at the same time. |
| 2 | Cafeteros | Physiological measures were cleaned and used to obtain derivative time series, such as phasic and tonic components of electrodermal activity. Using those signals, the team trained hybrid deep learning models consisting of convolutional and recurrent branches. Separate models were trained for valence and arousal ratings. |
| 3 | CARElab | Physiological signals were cleaned, split into windows, and used to derive descriptive features. Next, machine learning models were trained for narrowest possible context in each scenario using AutoGluon, a machine learning framework for automated model training and optimization. |
| 4 | UBA | Tree-based ensemble models (random forest and gradient-boosted decision trees) were trained to predict valence and arousal ratings in narrowest possible context, using features derived from preprocessed physiology as an input. |
| 5 | PSU | Physiological signals were cleaned and used to compute dynamic features. These features were used as an input for trained machine learning models (transformer models or ensembles of decision trees). |
| 6 | Northeastern | Physiological signals were cleaned and descriptive features were derived. Several classical machine learning models were trained, and the best model for each validation approach and affect dimension was chosen. Annotations were predicted every 1 second and later upsampled. |
| 7 | NYIT | Raw physiology was scaled and used as input for a custom version of FEDformer: a deep learning transformer-like architecture for time series forecasting. The same model predicted ratings of arousal and valence, which were later smoothed using a convolutional deep learning model. |
| 8 | Princeton | For across-time validation, signal windows with annotations centered relative to physiology were used as input for a LSTM model with information about subjects and videos. For other validation approach, the team used signal windows with annotations at the end, and regularized LSTM models working on physiology provided for challenge. Valence and arousal ratings were predicted simultaneously. |
| 9 | Queens | Preprocessed ECG signal (time-series) was used to train deep neural networks: convolutional and transformer-based architectures. Two approaches were explored: (1) training models from scratch, and (2) pretraining models on the entire training data and retraining them on narrower context. Out of tested approaches, transformer model trained from scratch on narrowest possible context achieved the best accuracy. |
| 10 | SAIL | Physiology was cleaned and resampled. Skin temperature signal was used in form of features, and the rest of signals were used as time series. A deep state-space S4 model and transfer learning was used for modeling, wherein layers for ECG signal processing were trained on a separate dataset. Different models were used to predict valence and arousal. |
| 11 | IITB | Physiology was cleaned and resampled. Data were split into 2 second windows with annotations at the end of each window. Separate decision trees were trained for predicting valence and arousal ratings. |
| 12 | VSL | Physiological signals were used as input to transformer-based models. Inside the model, physiology signals were normalized, scaled, and then processed by neural network layers. Models were trained to predict valence and arousal ratings at the same time. |