Recognition of Engagement from Electrodermal Activity Data Across Different Contexts

Leonardo Alchieri
leonardo.alchieri@usi.ch
Università della Svizzera italiana (USI)
Lugano-Viganello, Switzerland

Lidia Alecci
lidia.alecci@usi.ch
Università della Svizzera italiana (USI)
Lugano-Viganello, Switzerland

Nouran Abdalazim
nouran.abdalazim@usi.ch
Università della Svizzera italiana (USI)
Lugano-Viganello, Switzerland

Silvia Santini
silvia.santini@usi.ch
Università della Svizzera italiana (USI)
Lugano-Viganello, Switzerland

ABSTRACT

Engagement is a human experience relevant in multiple contexts, including classrooms, presentations and workplaces. Stemming from flow theory, engagement in these contexts has been studied using wearable devices, which can unobtrusively measure physiological changes, specifically Electrodermal Activity (EDA). However, researchers have not explored how EDA markers might be similar or different between various engagement scenarios, namely student, audience and workplace engagement. In this study, we investigated possible similarities through the use of three datasets containing EDA data and engagement self-report labels, collected in the wild in different settings using research-grade wrist-worn wearable devices. We analysed the correlation between hand-crafted EDA features and the engagement level and we leveraged a machine learning framework for engagement prediction. We found that similar features are correlated with the engagement level across the various settings. We also found that our machine learning model identified related markers as important across the three engagement contexts. Our results highlight that similarities are present in the EDA features between different engagement contexts, while also identifying possible dataset specific differences.

CCS CONCEPTS

• **Human-centered computing**; • **Computing methodologies**;

KEYWORDS

engagement; machine learning; wearable devices; contexts

ACM Reference Format:
Leonardo Alchieri, Lidia Alecci, Nouran Abdalazim, and Silvia Santini. 2023. Recognition of Engagement from Electrodermal Activity Data Across Different Contexts. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing (UbiComp/ISWC '23 Adjunct), October 8–12, 2023, Cancun, Quintana Roo, Mexico.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3594739.3610701

UbiComp/ISWC '23 Adjunct, October 8–12, 2023, Cancun, Quintana Roo, Mexico

ACM ISBN 979-8-4007-0200-6/23/10.
https://doi.org/10.1145/3594739.3610701

1 INTRODUCTION

Engagement, a multifaceted construct, is a human experience that can be found in various domains, including education, the workplace, and interpersonal communication. While engagement is defined differently in various domains, all concepts can find their roots in "flow" [12], a state of absorption, concentration and pleasure derived from compelling activities [44].

It is possible to leverage unobtrusive engagement tracking across various contexts: to improve academic achievements through targeted teaching [10, 19], to enhance workers' productivity and satisfaction by facilitating "flow" state experiences [32], and to refine professional performances and identify compatible audiences by gauging their engagement [46].

Recent development of wearable technology created new possibilities to perform engagement monitoring. Unobtrusive sensing of engagement levels has already been explored in education [16, 33, 48], workplace [14, 30, 39, 40], and audience-presenter contexts [22, 26, 38, 47], using distinct devices and physiological markers. Electrodermal Activity (EDA), due to its intimate association with autonomic nervous system responses [6], has been adopted as marker for the detection of flow and engagement in the various contexts aforementioned [14, 16, 22, 27].

All types of engagement mentioned, can find their roots in flow theory [3, 12]. Student engagement is defined by flow constructs [9, 10, 36, 42, 43]; audience engagement is linked to how our mind's "flow" during social interactions [11, 23, 25, 45]; and workplace engagement is usually defined directly as "flow" [4, 18]. However, these similarities have not yet been explored using wearable devices.

As such, our study made a novel contribution by comparing different engagement contexts, audience, workplace and student engagement, under a unified methodological framework. We investigated whether there are common EDA indicators of engagement by considering three distinct datasets, all containing EDA data from wearable devices and engagement self-reports in different contexts. To this end, we developed both a correlation and a machine learning framework to analyse how physiological markers interact with individuals' engagement level. We then tested our models' generalizability and adaptability in predicting engagement using six validation paradigms. The code used for this study is open sourced at: https://github.com/LeonardoAlchieri/EngagementPredictions. All three datasets are available to other researchers upon signing a data sharing agreement.

2 RELATED WORK

Di Lascio et al. [14] implemented a deep learning-based approach to recognize flow level in the workplace. Through Blood Volume Pulse (BVP) and EDA data collected from Empatica E4 devices in the wild, they managed to obtain a balanced accuracy of 70%. Instead, Rissler et al. [39] used, for a similar task, only EDA data from medical grade devices, in a lab setting, and a Support Vector Machine classifier [5], achieving 70% accuracy. Similarly, Rissler et al. [40] used Heart Rate Variability (HRV) data collected from chest-worn wearable devices and a Random Forest classifier, obtaining an accuracy of 70% from both in the wild and controlled conditions.

Di Lascio et al. [16] used EDA hand-crafted features and a machine learning classifier to predict students' emotional engagement during lectures, achieving a recall of 80%. Other studies [21, 33, 48] have investigated the connection between EDA markers and engagement during lectures.

Gashi et al. [22] showed a Dynamic Time Warping (DTW) method to recognize when the engagement level of an audience and a presenter is synchronized. Also, Röggla et al. [41] used a real-time EDA-based system to gauge audience engagement as part of an art installation. Similarly, Wang et al. [49] mapped audience's EDA signals and identified minute-per-minute engagement correlations.

There is a lack of work in analysing engagement across different settings and finding common patterns and markers, specifically from EDA data collected from wearable devices. As such, in this work we try to bridge this gap through the use of correlation analysis and a machine learning framework.

3 METHOD

3.1 Engagement Datasets

We use three distinct datasets, all obtained using wrist-worn Empatica E4 devices[1] for unobtrusive, continuous collection of EDA data. These datasets focus on different engagement contexts, i.e., student, workplace, and audience, collected through the use of different self-report questionnaires.

The **SEED** dataset [16] contains physiological data from 24 participants collected during nine lectures. Engagement self-reports, based on the "University Student Engagement Inventory" (USEI) questionnaire [31], were administered twice per lecture, yielding a computed score in $[1, 5]$ (floating-point). The **APSYNC** dataset [22] contains physiological data from 10 audience members across multiple presentations. A self-report questionnaire derived from Hassib et al. [26] gathered engagement data on a 7-point Likert scale. The **Workplace** dataset [14, 15] includes data from 14 academic workers performing various tasks over 28 days. Using the modified "Work-Related Flow Inventory" (WOLF) questionnaire [3, 13], administered after each work activity, the authors computed a flow score in $[1, 5]$.

3.2 Dataset Pre-processing

We filtered the signals utilizing a first-order Butterworth filter with a 0.4 Hz cutoff frequency, as [17]. Then, we decomposed the EDA signal into its tonic and phasic components [6], using the cvxEDA method [24]. We shall refer to these components as "phasic-EDA" and "tonic-EDA", while to the non-decomponsed signal as "mixed-EDA".

[1] https://www.empatica.com/en-gb/research/e4/

Table 1: Overview of hand-crafted EDA features extracted from 10-seconds segments, for all three engagement datasets.

Feature type	Feature list
time-domain	min, max, mean, std, dynamic range, slope, absolute slope, mean first derivative, std first derivative, number of peaks, peaks amplitude
wavelet-based	mean 1 Hz wavelet, std 1 Hz wavelet, mean means 2 Hz wavelet, std 2 Hz wavelet, mean 4 Hz wavelet, std 4 Hz wavelet
skin response	rise time, decay time

Following [15, 16, 22], we binarized the engagement labels into "low" and "high engagement". In the SEED and Workplace datasets, with a label scale of 1 to 5, we assigned "high engagement" for values above 3, as [15, 16]; while for APSYNC, which uses a scale of 1 to 10, the threshold was set to 5.

We then segmented each EDA component into non-overlapping windows of 10 seconds, following [16], for all three datasets. To each segmented window, the corresponding engagement label ("low engagement"/"high engagement") was assigned. We extracted a total of 13'985 windows for SEED, 1'951 for APSYNC and 41'890 for Workplace. The distribution of engagement labels is approximately 55% "low engagement" and 45% "high engagement" for all three datasets: no rebalancing was performed.

On each 10-second window, we then extracted 19 hand-crafted features per component (phasic, tonic and mixed-EDA), for a total of 57: 11 time-domain features, as [17], 6 wavelet-based features, as in [14], and 2 features that characterize the skin conductance response (SCR) [6]. In Table 1 we show an overview.

3.3 Feature Correlation with Engagement

We sought common patterns among the datasets via a correlation analysis between engagement labels and extracted features. The goal was to identify features potentially correlated with the engagement level across datasets. For this purpose, we utilized Spearman's rank correlation coefficient, measuring non-linear dependencies between the extracted features and engagement labels [29].

3.4 Machine Learning Classifiers

We employed 27 Machine Learning (ML) models to classifying the engagement level in the three contexts. With a similar methodology to [14, 16], each model took as input a set of hand-crafted EDA features, corresponding to a 10-second window, and then made a binary prediction: "low" or "high engagement".

We report only the results obtained by the Random Forest classifier [2], which achieved the highest balanced accuracy in most of our experiments, and was one of the best classifiers in similar studies [14, 16]. We used a Random Biased Guess predictor as baseline, which classifies engagement labels following the data distribution. All models were implemented in Python using the Scikit-Learning [37] and LazyPredict (https://github.com/shankarpandala/lazypredict) libraries.

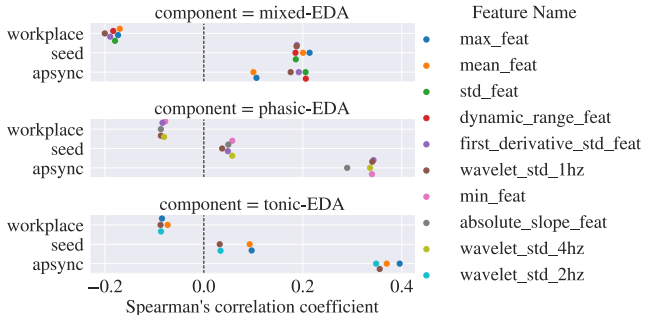

Figure 1: Spearman's ρ, across the three datasets. The feature represented are the top-2 highest correlated feature per dataset and per EDA component. Less than 6 points per comparison might be present, if two datasets share same features.

3.5 Machine Learning Evaluation Procedures

We employed various evaluation procedures on our machine learning task, to simulate real-world applicability, e.g., same user during the same day, unseen users or new days. We also tested how the models performed when trained on each dataset independently ("single-dataset training"), and using data from multiple datasets together ("multi-dataset training").

We used classic **5-fold cross-validation** to understand the models' ability to handle intra-user, intra-day data. Although it might not reflect real-world use, it still showcases the models' capabilities. We also used a **modified** version of **5-fold cross-validation**, following [14], where for each user, data from the same day was not concurrently present in both training and testing sets. Since the Workplace dataset has more data than the other two, for these validations we only used the "single-dataset training" paradigm.

We also leveraged **Leave Out Participant Out** (LOPO) and **Leave One Session Out** (LOSO) cross validation, to simulate testing on unseen users or days respectively. We also performed a **personalized LOSO** validation, where unique ML models were trained for each participant, providing insight into the influence of individual users' data. These three validations were run on both "single-dataset" and "multi-dataset training" paradigms.

To explore adaptability across engagement scenarios, we performed **Leave-One-Dataset-Out** (LODO) cross-validation, on the "multi-dataset training" paradigm only. This involved training the models on two datasets, while the third one was left out for testing.

Each procedure was computed with multiple seeds to account for stochastic phenomena, i.e., higher or lower performance due to initialization of the classifier or fold selection. We used balanced accuracy [8] as the evaluation metric, as [14], reported as average across all runs, with standard error.

3.6 Feature Importance Analysis

To assess if the three scenarios leveraged similar or different EDA features for engagement prediciton, we trained a Random Forest classifier using, for each dataset independently, all of the available data points. From it, we extracted the impurity-based feature importance [34]. We used this method since it highlights key features directly from the traine model. This allows for direct comparison of influential variables within each unique scenario. Future work could focus on other explainable methods, e.g., permutation importance [7], partial dependence plots or Shapley values [35].

4 RESULTS

4.1 Feature Correlation with Engagement

Figure 1 shows the engagement correlation, across EDA components. On the y-axis, we arrange EDA parts and datasets. The x-axis shows the correlation value. Each point shows how closely a feature relates to engagement level. We present the two most impactful features for each EDA part per dataset. The results show that correlation is similar, for the same feature, across scenario. For the Workplace dataset, features have negative correlation where the other two have positive correlation. Overall, all of our results are always lower, in absolute value, than 0.4, suggesting low correlation [20, 29, 50] with the engagement level. This result suggests that similar patterns might be present among the three contexts, but the low correlation values highlight the need for further inspection. As such, first we trained some Machine Learning models and then investigated which features were most relevant to predict the engagement label, in the three distinct scenarios.

4.2 Engagement Classification Results

In Table 2 we present the balanced accuracy results, with standard errors, for all of the validation paradigms, for both "single" and "multi-dataset training".

For **5-fold cross validation**, the Random Forest classifier reached a balanced accuracy above 80% when trained independently on all datasets, indicating successful engagement level recognition in all three scenarios when user and day data is shared between train and test sets. In **Leave Out Participant Out** (LOPO) cross validation, balanced accuracy surpasses the random baseline only with independent training on the APSYNC dataset, implying that physiological responses and perceived engagement levels can differ among participants in similar situations.

In the **Leave One Session Out** (LOSO) cross validation, performance for both "single" and "multi-dataset training" reaches approximately 60% balanced accuracy, hinting at inter-day differences possibly impacting the generalizability of work and audience engagement predictions. The **personalized Leave One Session Out** approach aims to address this issue, with noticeable performance enhancements observed across all contexts, even though always lower than when performing 5-fold cross validation. In the **modified 5-fold cross validation**, where data from the same day is never shared between training and testing folds for each user, performance is lower than the 5-fold cross validation but higher than Personalized LOSO only on the APSYNC dataset, hinting that mixing data from various users on different days in the training set might affect classifier performance. Despite implementing differences in pre-processing, segmentation, and feature extraction, the balanced accuracy on the Workplace dataset is around 60%, lower than the results from Di Lascio et al. [14], which however leveraged other features. Finally, results for the **Leave Out Dataset Out** (LODO) cross validation suggest that models trained on a different engagement scenarios fail to generalize accurately to others, with balanced accuracy for all datasets less than the random baseline.

Table 2: Balanced accuracy in % (with standard error) for the different validation paradigms on the Random Forest classifier. Results are presented for both "single-dataset" and "multi-dataset training". The hasterisk ∗ indicates results which satistically have higher balanced accuracy than the Random Biased Guess baseline.

Validation Paradigm	"Single-dataset training"			"Multi-dataset training"		
	Workplace	SEED	APSYNC	Workplace	SEED	APSYNC
5-fold	82.7(0.1)∗	80.5(0.4)∗	93.4(0.1)∗	/	/	/
LOPO	52.8(2.2)	44.7(5.1)	57.1(0.1)∗	49.5(0.1)	42.6(0.2)	42.2(0.2)
LOSO	60.0(0.2)∗	45.8(0.3)	57.4(0.2)∗	57.2(0.2)∗	41.5(0.2)	48.7(0.1)
Personalized LOSO	68.7(2.0)∗	78.7(9.6)	58.1(10.1)∗	/	/	/
Modified 5-fold	59.7(0.2)∗	47.6(0.6)	67.9(8.7)∗	/	/	/
LODO	/	/	/	40.5(0.1)	42.4(0.2)	42.9(0.6)

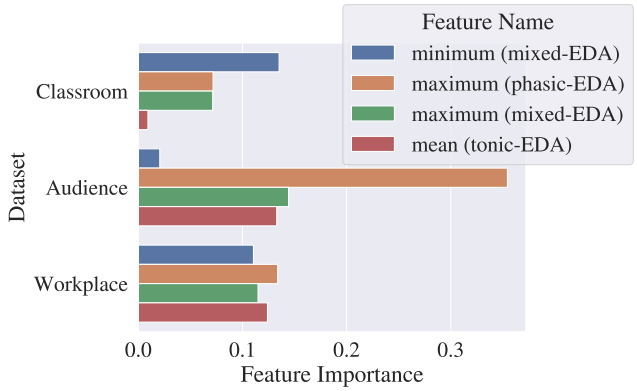

Figure 2: Displayed is the Feature Importance of the top-2 features per dataset based, when the Random Forest classifier is trained independently on three datasets. Shared top-2 features across datasets are reported only once, resulting in fewer than six unique features.

In conclusion, our validation paradigms suggest that over 80% balanced accuracy in engagement prediction is achievable only when same-day data from a single user is utilized for both training and validation, potentially due to data leakage [28]. When predicting on unseen points from an unknown user or day, the model's performance is often not different than the random baseline. We also highlighted that, unlike other scenarios where wearable devices are used, merging data from all three datasets doesn't improve engagement prediction [1, 51]. This is could be due to variations in the definition of engagement in the three contexts, different EDA signal markers or data imbalance, given the Workplace dataset's larger size compared to APSYNC and SEED (subsection 3.2).

4.3 Feature Importance

Figure 2 displays feature importance across three datasets, revealing similar significant features in each, despite some being dataset-specific. Most relate to operations like minimum, maximum, or average across different EDA components ("phasic", "tonic", "mixed"), indicating potential similarities across engagement tasks and the reliance of engagement prediction on common markers with some unique dataset variations.

5 CONCLUSIONS

In this work, we compared engagement across different contexts, namely audience, student and workplace engagement. We used Electrodermal Activity (EDA) datasets collected using Empatica E4 devices and engagement self-reports for the three settings. We performed the same pre-processing, decomposition, segmentation and extraction of hand-crafted EDA features. We then correlated the EDA features with the engagement level. Finally, we created a machine learning pipeline to predict the engagement level of individuals and to analyse which hand-crafted features might be leveraged the most in the classification.

Our results highlight similarities between audience, student and workplace engagement. We showed that the hand-crafted features with highest correlation are similar across datasets, hinting to common physiological responses. To expand on this, our feature importance analysis found that similar features were also leveraged in the three scenarios.

Further work is nonetheless necessary, especially to improve engagement prediction. Future analysis should also investigate how to leverage data from multiple contexts and analyse how features intertwine with each other.

In conclusion, our work tried to bridge the gap in engagement recognition knowledge by highlighting the presence of similarities between various scenarios, i.e., audience, student and workplace engagement, using correlation analysis and a machine learning classification task. We pose the basis for further investigation into how physiological signals from wearable devices might be leveraged across different engagement contexts.

ACKNOWLEDGMENTS

This work is partially supported by the Swiss National Science Foundation (SNSF) through the grant 205121_197242 for the project "PROSELF: Semi-automated Self-Tracking Systems to Improve Personal Productivity".

REFERENCES

[1] Daniel A Adler, Fei Wang, David C Mohr, and Tanzeem Choudhury. 2022. Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *Plos one* 17, 4 (2022), e0266516.

[2] Taiwo Oladipupo Ayodele. 2010. Types of machine learning algorithms. *New advances in machine learning* 3 (2010), 19–48.

[3] Arnold B Bakker. 2005. Flow among music teachers and their students: The crossover of peak experiences. *Journal of vocational behavior* 66, 1 (2005), 26–44.

[4] Arnold B Bakker. 2008. The work-related flow inventory: Construction and initial validation of the WOLF. *Journal of vocational behavior* 72, 3 (2008), 400–414.

[5] Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*. Vol. 4. Springer.

[6] Wolfram Boucsein. 2012. *Electrodermal activity*. Springer Science & Business Media.

[7] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[8] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. 2010. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*. IEEE, 3121–3124.

[9] Jere Brophy. 1983. Conceptualizing student motivation. *Educational psychologist* 18, 3 (1983), 200–215.

[10] Sandra Christenson, Amy L Reschly, Cathy Wylie, et al. 2012. *Handbook of research on student engagement*. Vol. 840. Springer.

[11] Karen E Cooper. 2009. *Go with the flow: Examining the effects of engagement using flow theory and its relationship to achievement and performance in the 3-dimensional virtual learning environment of Second Life*. University of Central Florida.

[12] Mihaly Csikszentmihalyi. 1990. Flow: the psychology of optimal experience. *Harper Perennial modern classics* (1990).

[13] Reeshad S Dalal, Holly Lam, Howard M Weiss, Eric R Welch, and Charles L Hulin. 2009. A within-person approach to work behavior and performance: Concurrent and lagged citizenship-counterproductivity associations, and dynamic relationships with affect and overall job performance. *Academy of Management Journal* 52, 5 (2009), 1051–1066.

[14] Elena Di Lascio, Shkurta Gashi, Maike E Debus, and Silvia Santini. 2021. Automatic Recognition of Flow During Work Activities Using Context and Physiological Signals. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.

[15] Elena Di Lascio, Shkurta Gashi, Juan Sebastian Hidalgo, Beatrice Nale, Maike E Debus, and Silvia Santini. 2020. A multi-sensor approach to automatically recognize breaks and work activities of knowledge workers in academia. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–20.

[16] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2018. Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–21.

[17] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2019. Laughter recognition using non-invasive wearable devices. In *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*. 262–271.

[18] Stefan Ed Engeser. 2012. *Advances in flow research*. Springer Science+ Business Media.

[19] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. 2004. School engagement: Potential of the concept, state of the evidence. *Review of educational research* 74, 1 (2004), 59–109.

[20] David Freedman, Robert Pisani, and Roger Purves. 2007. *Statistics*.

[21] Enqi Fu, Matias Laporte, Cindy Guerrero Toro, Martin Gjoreski, and Marc Langheinrich. 2022. Affect and Learning in the LAUREATE Dataset. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*. 168–172.

[22] Shkurta Gashi, Elena Di Lascio, and Silvia Santini. 2019. Using unobtrusive wearable sensors to measure the physiological synchrony between presenters and audience members. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 1 (2019), 1–19.

[23] Lakshmi Goel, Norman Johnson, Iris Junglas, and Blake Ives. 2013. Predicting users' return to virtual worlds: a social perspective. *Information Systems Journal* 23, 1 (2013), 35–63.

[24] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. 2015. cvxEDA: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering* 63, 4 (2015), 797–804.

[25] Austin M Grinberg, Jesus Serrano Careaga, Matthias R Mehl, and Mary-Frances O'Connor. 2014. Social engagement and user immersion in a socially based virtual world. *Computers in Human Behavior* 36 (2014), 479–486.

[26] Mariam Hassib, Stefan Schneegass, Philipp Eiglsperger, Niels Henze, Albrecht Schmidt, and Florian Alt. 2017. EngageMeter: A system for implicit audience engagement sensing using electroencephalography. In *Proceedings of the 2017 Chi conference on human factors in computing systems*. 5114–5119.

[27] Javier Hernandez, Zicheng Liu, Geoff Hulten, Dave DeBarr, Kyle Krum, and Zhengyou Zhang. 2013. Measuring the engagement level of TV viewers. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 1–7.

[28] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. 2012. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6, 4 (2012), 1–21.

[29] Maurice George Kendall et al. 1946. The advanced theory of statistics. *The advanced theory of statistics*. 2nd Ed (1946).

[30] Matthew Lee. 2020. Detecting affective flow states of knowledge workers using physiological sensors. *arXiv preprint arXiv:2006.10635* (2020).

[31] João Maroco, Ana Lúcia Maroco, Juliana Alvares Duarte Bonini Campos, and Jennifer A Fredricks. 2016. University student's engagement: development of the University Student Engagement Inventory (USEI). *Psicologia: Reflexão e Crítica* 29 (2016).

[32] Andrew J Martin. 2005. The role of positive psychology in enhancing satisfaction, motivation, and productivity in the workplace. *Journal of Organizational Behavior Management* 24, 1-2 (2005), 113–133.

[33] Karen S McNeal, Jacob M Spry, Ritayan Mitra, and Jamie L Tipton. 2014. Measuring student engagement, knowledge, and perceptions of climate change in an introductory environmental geology course. *Journal of Geoscience Education* 62, 4 (2014), 655–667.

[34] Bjoern H Menze, B Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A Hamprecht. 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics* 10 (2009), 1–16.

[35] Christoph Molnar. 2020. *Interpretable machine learning*. Lulu. com.

[36] Giovanni B Moneta and Mihaly Csikszentmihalyi. 1996. The effect of perceived challenges and skills on the quality of subjective experience. *Journal of personality* 64, 2 (1996), 275–310.

[37] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[38] Rosalind W Picard and Jocelyn Scheirer. 2001. The galvactivator: A glove that senses and communicates skin conductivity. In *Proceedings 9th Int. Conf. on HCI*.

[39] Raphael Rissler, Mario Nadj, Maximilian Xiling Li, Michael Thomas Knierim, and Alexander Maedche. 2018. Got flow? Using machine learning on physiological data to classify flow. In *Extended abstracts of the 2018 CHI conference on human factors in computing systems*. 1–6.

[40] Raphael Rissler, Mario Nadj, Maximilian Xiling Li, Nico Loewe, Michael T Knierim, and Alexander Maedche. 2020. To be or not to be in flow at work: physiological classification of flow using machine learning. *IEEE transactions on affective computing* (2020).

[41] Thomas Röggla, Chen Wang, Lilia Perez Romero, Jack Jansen, and Pablo Cesar. 2017. Tangible air: an interactive installation for visualising audience engagement. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*. 263–265.

[42] David J Shernof, Erik A Ruzek, Alexander J Sannella, Roberta Y Schorr, Lina Sanchez-Wall, and Denise M Bressler. 2017. Student engagement as a general factor of classroom experience: Associations with student practices and educational outcomes in a university gateway course. *Frontiers in Psychology* 8 (2017), 994.

[43] David J Shernoff. 2010. *The experience of student engagement in high school classrooms: Influences and effects on long-term outcomes*. LAP Lambert Academic Publishing.

[44] David J Shernoff. 2013. Optimal learning environments to promote student engagement. (2013).

[45] Patricia A Thomas. 2011. Gender, social engagement, and limitations in late life. *Social science & medicine* 73, 9 (2011), 1428–1435.

[46] Ben Walmsley. 2019. *Audience engagement in the performing arts: A critical analysis*. Springer.

[47] Chen Wang and Pablo Cesar. 2015. Measuring Audience Responses of Video Advertisements using Physiological Sensors.. In *ImmersiveME@ ACM Multimedia*. 37–40.

[48] Chen Wang and Pablo Cesar. 2015. Physiological Measurement on Students' Engagement in a Distributed Learning Environment. *PhyCS* 10 (2015), 0005229101490156.

[49] Chen Wang, Erik N Geelhoed, Phil P Stenton, and Pablo Cesar. 2014. Sensing a live audience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1909–1912.

[50] Robert S Witte and John S Witte. 2017. *Statistics*. John Wiley & Sons.

[51] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S Kuehn, Jeremy F Huckins, Margaret E Morris, et al. 2023. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–34.