Predicting Psychological States from Computer Activity Data

Ana Poklukar
ap3956@student.uni-lj.si
Jožef Stefan Institute
Ljubljana, Slovenia

Gabrijel Okorn
gabrijel.okorn@gmail.com
Jožef Stefan Institute
Ljubljana, Slovenia

Leon Costa
leon.costa@usi.ch
Università della Svizzera Italiana
Lugano, Switzerland

Gašper Slapničar
gasper.slapnicar@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia

Mitja Luštrek
mitja.lustrek@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia

Zoja Anžur
zoja.anzur@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia

Abstract

A person's psychological state affects both their well-being and how well they work, but measuring it usually means interrupting them with a questionnaire. We investigate whether seven psychological states (emotional valence and arousal, fatigue, stress, productivity, engagement, and a well-being composite) can instead be predicted from passively logged computer activity. Using ActivityWatch, we collected data from 17 participants and engineered 16 behavioural features, then evaluated six models across three splitting strategies (including within-user and leave-one-subject-out) and three ways of framing the prediction task (regression and two classification variants). Gains over a dummy baseline vary considerably between states: productivity and engagement are the easiest to predict, while valence is essentially unpredictable from this signal alone, based on our data. Personalized approaches, especially within-user models, consistently outperform cross-user ones, suggesting psychological states manifest differently across individuals. Computer activity is shown to be a weak but informative predictor of psychological state.

Keywords

psychological state prediction, computer activity data, workplace well-being, machine learning

1 Introduction

A person's psychological state strongly influences their well-being and work performance [13]. The ability to automatically recognize psychological states could support future applications in healthcare, workplace well-being, and adaptive user interfaces. Such states are typically assessed with self-report questionnaires, which interrupt the user and cannot be administered continuously. Passive sensing with wearable devices avoids this interruption, but requires dedicated hardware.

Computer activity, in contrast, is a continuously available behavioural signal that requires no extra hardware and does not disturb the user. The goal of this study was therefore to investigate whether psychological states can be predicted from computer activity data alone.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.skui.70

2 Related Work

Prior work has linked computer use with psychological states: stress from keystrokes and language [15], emotions from typing patterns [10], focus, boredom and stress from digital activity [13], and mood from combined computer and wearable data [12]. As reviewed by Anžur et al. [1], other sensors such as microphones, cameras and smartphones can also monitor psychological states. Despite this range of methods, prediction from unobtrusive, non-contact sensors remains challenging, in part because these states manifest differently across individuals, a source of variability that [15, 10, 13, 12] do not explicitly model.

We present a framework for predicting psychological states (valence, arousal, fatigue, stress, productivity, engagement, and well-being) from application usage, and examine whether models generalize across individuals or need to be personalized to each one.

3 Data

3.1 Data Collection

Data were collected from 17 participants during their regular working days between July 2025 and March 2026. Participation was spread across this period rather than simultaneous, with each participant contributing between 13 and 30 working days. The study recorded several modalities; this work focuses only on computer activity and the accompanying self-reports of psychological states.

Self-reports were collected several times per day using ecological momentary assessment (EMA) [14]: nine questions about momentary psychological state, all on a 7-point scale, yielding roughly 950 responses per question. Emotional valence, physical comfort and workplace interactions were rated −3 to +3; arousal, fatigue, immersion, enthusiasm, stress and productivity were rated 0 to 6. Answers were given via Stream Deck, a computer accessory with 15 programmable buttons, which kept interruptions minimal.

Computer activity was recorded with ActivityWatch [2], an automated time-tracking application running in the background, logging the active application, window title, and periods of keyboard and mouse inactivity.

3.2 Data Preprocessing

The raw ActivityWatch [2] events (application usage of varying duration) were segmented into overlapping 3-second windows (25% overlap). Since meaningful usage patterns occur on a longer

time scale, each window was described using a full one-minute interval around it (extended in both directions), not the 3 seconds alone.

The logged application name and window title are too specific and unstructured to use directly, so we mapped events onto 6 categories and 15 subcategories (e.g., *work*, *programming*) via keyword regular expressions, which served as the basis for the hand-crafted features describing behaviour in each window.

Self-reported answers were converted to a common scale (originally on different ranges), with fatigue and stress reversed so that higher values consistently indicated a more favourable state. Besides predicting each psychological state on its own, we formed two composite labels: engagement, the average of the complementary enthusiasm and immersion items; and well-being, combining valence and engagement, chosen based on Cronbach's alpha [9], since excluding stress and fatigue improved internal consistency.

Finally, we labelled windows with the answers given by participants using a larger interval than the one-minute feature context above: since a self-report reflects the preceding period rather than one instant, answers at a prompt were assigned to all windows within a fixed interval before it (15 or 30 minutes, treated as separate experiments); windows outside any interval were discarded. Individual labels have some missing responses (1.5–6.4%); each target was trained only on rows where it was answered, so a missing state does not affect predictions for others.

4 Methodology

This section describes the features used to represent each window, how the prediction task and evaluation protocol were set up, and which models and metrics were used.

Features. Each instance is one 3-second window (as defined in Section 3.2), labelled via its associated prompt response, described by 16 numeric features computed over its one-minute context window. Five features describe activity type: development, communication, research, personal-distraction and AI-assistant ratios, each the share of active time spent on that activity. Three describe computer use: input-active ratio (share of the minute with keyboard or mouse input), app-switch rate, and mean focus-block duration (average time in one activity category before switching). Four use the previous minute: previous-minute writing ratio (writing and office work), and change in personal-distraction ratio, development ratio and app-switch rate from that minute. The last four describe fragmentation: focus fragmentation (how activity is spread across applications and how often it switches), dominance gap (difference between the two most-used types), and median and standard deviation of segment length (typical duration and variability of time a window stays active).

Prediction Task Formulation. We compared three ways of framing the machine learning prediction task: Regression on the native 0–6 scale, Fixed classification (binarizing at the scale's midpoint, 3, for everyone), and Personalized classification (binarizing at each subject's own train-only median, to avoid leakage; only defined for within-user, since it needs training data from a single subject). Several targets have a population mean above the midpoint, so Fixed classification produces more imbalanced, sometimes degenerate, class splits than Personalized does.

Evaluation Protocol. We used three splitting strategies, each giving one fold per subject (17 folds total), and never shuffled windows before splitting, since neighbouring windows are seconds apart and near-identical, and a random split would leak duplicates from test into train. Leave one subject out (LOSO) trains on the other 16 subjects and tests on the held-out one; Within-User trains on that subject's earliest 70% and tests on their most recent 30%; LOSO-Personalized combines both, training on the other 16 subjects plus the held-out subject's earliest 30% and testing on their remaining 70%. Folds that end up single-class after binarizing are dropped rather than scored, which happens more with Fixed-threshold folds than Personalized ones (19% vs. 9%), matching the skew above.

Models. We used six models: Random Forest [3], LightGBM [11], XGBoost [5], SVM with an RBF kernel [6], KNN [7], and Logistic Regression (classification only) [8], plus a dummy baseline predicting the majority class for classification and the training mean for regression. We used library defaults for hyperparameters, except fixing 200 trees for the tree-based models, raising Logistic Regression's iteration cap to 1,000 for convergence, and fixing the random seed. SVM, KNN and Logistic Regression are scale-sensitive, so their inputs were standardized using the mean and standard deviation of the training fold only, to avoid leakage. SVM's solver was also capped at 10,000 iterations, and on some folds it hit this cap without converging, so its results warrant more caution.

Metrics. We evaluated regression with R^2 (≤ 1, no lower bound), chosen over an error metric like MAE or RMSE because it is normalized and comparable across our seven targets despite their different scales and variances. For classification we used balanced accuracy [4], chosen over raw accuracy since our classes are imbalanced: on balanced accuracy, a majority-class predictor always scores exactly $1/n_{\text{classes}}$, giving a fixed, interpretable dummy reference. In addition to these two scores for our models, we recorded dummy score, gain (model score minus dummy score on the same fold), accuracy, F1, and the confusion matrix, but only the main score and gain are compared across the full grid; the rest are shown only for the best-performing target.

5 Results

Table 1 shows Within-User usually performs best, since personalization calibrates the model to the individual rather than relying on patterns pooled across everyone. Gains also vary a lot between experiments, so the averages below should be read with caution, as Table 5 and Figure 1 show.

Averaged across all targets, models and strategies, Regression gain is negative (-0.96), while both classification variants are positive, with Personalized (+0.03) beating Fixed (+0.01) as the imbalance argument from Section 4 would predict.

Table 2 shows the models perform similarly for classification (gain from +0.013 to +0.020), but not for Regression: tree models stay in a similar range (-0.24 to -0.38), while SVM and KNN fall much further behind (-1.82 and -2.04). This lines up with the convergence and scale-sensitivity issues we noted in Section 4.

Table 3 makes this concrete: Within-User's classification gain (+0.034) is more than ten times that of LOSO and LOSO-Personalized (both +0.002), reinforcing that personalization, not model choice or threshold, is what mainly helps here.

Table 4 shows the 15-minute window winning for 4 of the seven targets, usually by a small margin. Productivity, engagement and well-being are the exception, doing slightly better with the longer 30-minute window. So there is no single best window size, but 15 minutes is the safer default.

Table 1: Gain (score minus dummy) for every target, strategy and task, by model. Grey rows are each target's best classification result.

Target	Strategy	Task	RF	LightGBM	XGBoost	SVM	KNN	LogReg	Dummy
	LOSO	Reg.	-0.23	-0.07	-0.14	-2.65	-0.99	–	-0.35
	LOSO	**Fixed**	**0.00**	**0.00**	**0.00**	**+0.01**	**0.00**	**0.00**	**0.50**
	LP	Reg.	-0.25	-0.03	-0.12	-4.34	-1.24	–	-0.57
valence	LP	Fixed	0.00	0.00	0.00	0.00	0.00	0.00	0.50
	WU	Reg.	-1.10	-1.01	-1.14	-0.72	-5.67	–	-0.24
	WU	Fixed	-0.02	0.00	-0.01	-0.02	-0.01	-0.01	0.50
	WU	Pers.	0.00	0.00	-0.01	+0.01	+0.01	-0.01	0.50
	LOSO	Reg.	-0.35	-0.15	-0.24	-1.98	-1.26	–	-1.16
	LOSO	Fixed	0.00	0.00	-0.01	-0.01	-0.03	0.00	0.50
	LP	Reg.	-0.28	0.00	-0.08	-4.25	-1.56	–	-3.04
arousal	LP	Fixed	0.00	0.00	-0.01	+0.01	-0.05	0.00	0.50
	WU	Reg.	-0.60	-0.52	-0.67	-0.74	-2.82	–	-0.45
	WU	Fixed	+0.04	+0.05	+0.04	+0.05	+0.03	+0.06	0.50
	WU	**Pers.**	**+0.06**	**+0.07**	**+0.06**	**+0.07**	**+0.05**	**+0.08**	**0.50**
	LOSO	Reg.	-0.44	-0.27	-0.34	-1.66	-0.71	–	-0.67
	LOSO	Fixed	+0.01	+0.01	+0.01	+0.02	+0.01	0.00	0.50
	LP	Reg.	-0.32	-0.09	-0.18	-1.65	-0.58	–	-0.52
fatigue	LP	Fixed	+0.01	+0.01	0.00	-0.01	+0.01	0.00	0.50
	WU	Reg.	-0.62	-0.57	-0.73	-0.79	-2.62	–	-0.46
	WU	Fixed	0.00	0.00	0.00	-0.01	-0.03	+0.01	0.50
	WU	**Pers.**	**+0.04**	**+0.04**	**+0.04**	**+0.03**	**-0.01**	**+0.05**	**0.50**
	LOSO	Reg.	-0.33	-0.14	-0.23	-1.59	-0.58	–	-0.84
	LOSO	Fixed	0.00	0.00	0.00	+0.02	0.00	0.00	0.50
	LP	Reg.	-0.59	-0.36	-0.41	-3.20	-1.02	–	-1.45
stress	LP	Fixed	0.00	0.00	0.00	+0.01	0.00	0.00	0.50
	WU	Reg.	-1.01	-0.92	-1.05	-1.29	-4.13	–	-0.75
	WU	**Fixed**	**+0.04**	**+0.04**	**+0.04**	**+0.05**	**+0.03**	**+0.05**	**0.50**
	WU	Pers.	0.00	0.00	0.00	+0.01	0.00	0.00	0.50
	LOSO	Reg.	-0.16	-0.02	-0.09	-1.25	-1.25	–	-0.39
	LOSO	Fixed	0.00	0.00	0.00	-0.02	+0.09	0.00	0.50
	LP	Reg.	-0.08	+0.06	0.00	-1.23	-1.20	–	-0.50
productivity	LP	Fixed	0.00	0.00	+0.01	-0.06	+0.11	-0.01	0.50
	WU	Reg.	-0.27	-0.26	-0.32	-0.33	-3.73	–	-0.23
	WU	Fixed	+0.07	+0.08	+0.07	+0.07	0.00	+0.08	0.50
	WU	**Pers.**	**+0.08**	**+0.09**	**+0.09**	**+0.07**	**-0.02**	**+0.08**	**0.50**
	LOSO	Reg.	-0.16	0.00	-0.09	-2.67	-1.95	–	-0.55
	LOSO	Fixed	0.00	0.00	0.00	-0.04	+0.06	0.00	0.50
	LP	Reg.	-0.15	+0.02	-0.06	-3.93	-2.42	–	-0.93
engagement	LP	Fixed	0.00	0.00	0.00	-0.01	+0.06	0.00	0.50
	WU	Reg.	-0.15	-0.14	-0.18	-0.22	-2.24	–	-0.35
	WU	**Fixed**	**+0.09**	**+0.09**	**+0.09**	**+0.08**	**+0.01**	**+0.07**	**0.50**
	WU	Pers.	+0.04	+0.05	+0.04	+0.03	+0.01	+0.03	0.50
	LOSO	Reg.	-0.19	-0.01	-0.09	-1.42	-0.90	–	-0.60
	LOSO	Fixed	0.00	0.00	0.00	-0.03	+0.01	0.00	0.50
	LP	Reg.	-0.15	+0.04	-0.07	-1.81	-1.01	–	-1.22
well-being	LP	Fixed	0.00	0.00	0.00	-0.01	+0.02	0.00	0.50
	WU	Reg.	-0.59	-0.55	-0.64	-0.47	-4.98	–	-0.24
	WU	**Fixed**	**+0.03**	**+0.03**	**+0.03**	**+0.03**	**+0.03**	**+0.02**	**0.50**
	WU	Pers.	+0.02	+0.03	+0.02	+0.03	+0.02	+0.02	0.50

Note: LP = LOSO-Personalized, WU = Within-User, Reg. = Regression, Pers. = Personalized. Reg. rows use R^2, Fixed/Pers. rows use balanced accuracy, so they are not on the same scale. LogReg only does classification. Dummy is its raw score, not gain.

Table 2: Average gain by model, split by task.

Model	Cls. Gain	Reg. Gain
RF	+0.019	-0.38
LightGBM	+0.020	-0.24
XGBoost	+0.018	-0.33
SVM	+0.013	-1.82
KNN	+0.016	-2.04
LogReg	+0.018	–

Note: Classification uses balanced accuracy, Regression uses R^2.

Table 3: Average gain by splitting strategy, across all targets and models.

Strategy	Cls. Gain	Reg. Gain
LOSO	+0.002	-0.70
LP	+0.002	-0.93
WU	**+0.034**	-1.25

Note: LP = LOSO-Personalized, WU = Within-User. Classification uses balanced accuracy, Regression uses R^2.

Table 5 zooms in on *productivity*, the target with the highest average classification gain. Accuracy looks high in places (up to 0.748 for Within-User Fixed), but this is largely because the classes are imbalanced: balanced accuracy for the same rows sits much lower, around 0.58, which is closer to the real picture. Figure 1 shows why averages need this caution: gain for individual subjects ranges from clearly negative to well above average, so a single mean hides a lot of person-to-person variation.

Table 4: 15-minute vs. 30-minute lookback window, best classification gain per target (best strategy/binarize/model combination for that target).

Target	15 min	30 min
valence	**+0.014**	0.005
arousal	**+0.080**	+0.066
fatigue	**+0.050**	+0.023
stress	**+0.049**	+0.027
productivity	+0.114	**+0.124**
engagement	+0.088	**+0.094**
well-being	+0.033	**+0.052**

Table 5: Extended metrics for *productivity* (highest average classification gain of the seven targets).

Strat.	Task	Model	Acc.	Bal. Acc.	F1	Gain
LOSO	Fixed	KNN	0.586	0.587	0.611	+0.087
LP	Fixed	KNN	0.593	0.614	0.621	+0.114
WU	Fixed	LogReg	0.748	0.580	0.722	+0.080
WU	Pers.	LightGBM	0.688	0.589	0.668	+0.089

Note: each row shows the best single model for that strategy/task setup, which can differ from Table 1's highlighted row, since that reflects the best average across all models rather than any single model's peak. LP = LOSO-Personalized, WU = Within-User, Pers. = Personalized, Bal. Acc. = balanced accuracy.

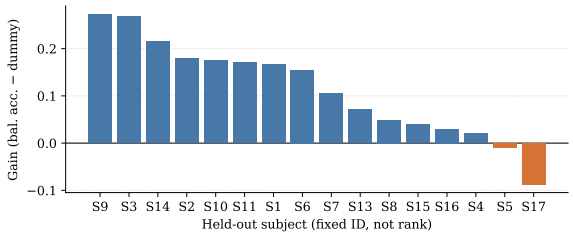


Figure 1: Gain by held-out subject for *productivity*.

Note: LOSO-Personalized, Fixed threshold, KNN (the best combination from Table 5). S12 is missing because its fold had only one class.

The confusion matrix in Figure 2 shows the same model is noticeably better at catching low-productivity windows (66% correctly classified) than high-productivity ones (56%). So even for our best-performing target, the model is still missing almost half of the actual high-productivity moments.

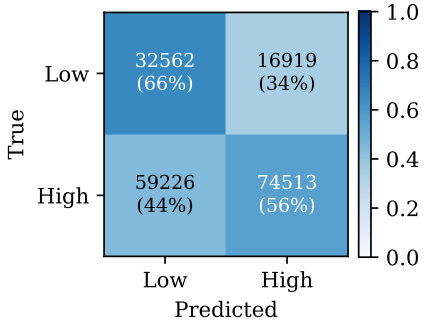


Figure 2: Confusion matrix for *productivity*, LOSO-Personalized, Fixed threshold, KNN, summed over all 16 folds. Shows raw counts and row percentages.

6 Conclusion

Using ActivityWatch [2], we predicted seven psychological states from 17 participants' computer activity data, using multiple evaluation strategies, prediction tasks, and machine learning algorithms. Our findings show substantial differences between states in prediction performance, indicating that psychological states are complex and need to be modelled individually. More personalized approaches performed better (best results with the Within-User approach), pointing to substantial individual differences in how these states manifest in computer use. Several targets also produce imbalanced, sometimes degenerate class splits under fixed thresholding; future work could test whether correction techniques or resampling help beyond what balanced accuracy already accounts for.

Acknowledgements

We thank our 17 study participants for their time and continued participation throughout the data collection period. This work was funded by the TRUST-ME project, jointly supported by the Swiss National Science Foundation (SNSF) (grant agreement 205121L_214991) and the Slovenian Agency of Research and Innovation (ARIS) (grant agreement N1-0319).

References

[1] Zoja Anžur et al. 2025. A review of methods for unobtrusive measurement of work-related well-being. *Machine Learning and Knowledge Extraction*, 7, 3, 62. doi:10.3390/make7030062.

[2] [SW] Erik Bjäreholt and Johan Bjäreholt, ActivityWatch version 0.13.1, June 2024. doi:10.5281/zenodo.4957165, URL: https://github.com/ActivityWatch/activitywatch.

[3] Leo Breiman. 2001. Random forests. *Machine Learning*, 45, 1, 5–32. doi:10.1023/A:1010933404324.

[4] 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. doi:10.1109/ICPR.2010.764.

[5] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '16). Association for Computing Machinery, 785–794. doi:10.1145/2939672.2939785.

[6] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20, 3, 273–297. doi:10.1007/BF00994018.

[7] Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13, 1, 21–27. doi:10.1109/TIT.1967.1053964.

[8] David R. Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20, 2, 215–232. doi:10.1111/j.2517-6161.1958.tb00292.x.

[9] Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 3, 297–334. doi:10.1007/BF02310555.

[10] Clayton Epp, Michael Lippold, and Regan L. Mandryk. 2011. Identifying emotional states using keystroke dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11). Association for Computing Machinery, 715–724. doi:10.1145/1978942.1979046.

[11] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: a highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 3146–3154.

[12] Gloria Mark, Mary Czerwinski, Shamsi T. Iqbal, and Paul Johns. 2016. Workplace indicators of mood: behavioral and cognitive correlates of mood among information workers. In *Proceedings of the 6th International Conference on Digital Health* (DH '16). Association for Computing Machinery, 29–36. doi:10.1145/2896338.2896360.

[13] Gloria Mark, Shamsi T. Iqbal, Mary Czerwinski, and Paul Johns. 2014. Bored Mondays and focused afternoons: the rhythm of attention and online activity in the workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '14). Association for Computing Machinery, 3025–3034. doi:10.1145/2556288.2557204.

[14] Saul Shiffman, Arthur A. Stone, and Michael R. Hufford. 2008. Ecological momentary assessment. *Annual Review of Clinical Psychology*, 4, 1, 1–32. doi:10.1146/annurev.clinpsy.3.022806.091415.

[15] Lisa M. Vizer, Lina Zhou, and Andrew Sears. 2009. Automated stress detection using keystroke and linguistic features: an exploratory study. *International Journal of Human-Computer Studies*, 67, 10, 870–886. doi:10.1016/j.ijhcs.2009.07.005.