A random split scores a stress classifier perfect on data with nothing to learn

A preregistered test of person-level generalisation on the WESAD chest recordings of 15 people, with planted data, a block-shuffle control and a felt-stress check

Zain Dana Harper

Version 1.0, 9 November 2026. Finding F320 in the findings ledger. Signal (tier B).

Abstract

Wearable stress detectors are often scored by cutting recordings into overlapping windows and splitting the windows at random into training and test sets. We tested what that split measures. On planted data built on the same windows and timing as the WESAD recordings, where each person and each condition block carries its own random offsets and nothing carries from one person to the next, a random split over windows scored an AUC of 1.000 and a held-out-person split scored 0.469. On the real chest recordings of 15 people (2,122 windows of 60 s, 30 percent of them stress), the random forest held across people at a held-out-person AUC of 0.966, against 1.000 under the random split. The random split inflated the AUC by 0.033 for the random forest and 0.086 for logistic regression. ECG, EDA and skin temperature together reached 0.97 for a held-out person, and acceleration with EMG reached 0.79. Across the 15 people, the rise in reported anxiety did not follow the size of the bodily change: Spearman rho 0.03 with EDA, 0.02 with heart rate and -0.45 with breathing rate (p 0.09). The first run failed its own planted check because we set an unreachable bar; a second preregistered run repaired the plant and passed.

1 What this means for a reader

If a paper reports a near-perfect stress classifier, look for how it split the data. When neighbouring windows share most of their samples and land on both sides of a random split, the model can score perfectly by recognising the recording, with no stress signal at all. In our test that happened exactly. On the real data the classifier did hold for new people, which is good news for this one laboratory protocol, but only a split that holds out whole people could show it.

2 Introduction

WESAD, the Wearable Stress and Affect Detection dataset of Schmidt and colleagues [1], records 15 people with a chest sensor and a wrist sensor during a baseline, amusing film clips and the Trier Social Stress Test, in which a person gives a short public talk and then does mental arithmetic aloud before an evaluating panel. The dataset paper reports classifiers scored with every person held out in turn, and WESAD has since become a common benchmark.

Many later studies score their models with a random split over windows: the windows are assigned to training and test sets at random, so windows from the same person and the same recording block fall on both sides. The windows are usually overlapping windows, fixed-length stretches cut at a step shorter than their length. Here a 60 s window is cut every 15 s, so neighbours share 75 percent of their samples.

The clinical machine-learning literature has argued over this choice. Saeb and colleagues [2] showed on wearable and phone sensor data that record-wise and subject-wise cross-validation can give very different error rates, and traced the gap to the classifier learning to recognise people. Little and colleagues [3] replied that

subject-wise splits can also mislead in some settings. Chaibub Neto and colleagues [4] built a permutation test for identity confounding and, on mobile health data, found strong evidence of it under record-wise splits. Our searches on 9 November 2026 also returned recent WESAD studies and open audits that make the same point about random splits; we read those at the level of search summaries and do not cite their numbers.

This paper adds a known answer. We build data in which nothing can carry from one person to another, run it through the same windows, features and models as the real data, and measure what the random split reports. Then we ask three questions of the real data: whether the classifier holds for a person it has not seen, which sensors carry the signal, and whether the size of a person's bodily response follows the size of the stress they report. On the last question, Campbell and Ehlert [5] reviewed acute psychosocial stress studies and asked whether emotional and physiological responses correspond.

3 Preregistration and design

The design was written as a preregistration, a dated plan fixed and hash-locked before any result is seen, on 31 October 2026 at 22:37 UTC. Window features were extracted and fixed before any label was compared with them. The decision it informs was stated in the file: how much to discount a wearable stress classifier reported under a within-person random split, and whether its label reads stress or a protocol.

The locked file names the data, the windows and features, four splits, two models, the sensor groups, the felt-stress test, a planted check (data built so the right answer is known, run before the real data), a control (a version of the data with the structure under test removed), forecasts and kill conditions, rules that decide against the claim with no room for judgement after the run.

Kill conditions, without discretion:

- If the planted data with nothing to learn fail (random split under 0.8 or held-out person above 0.6), or if the planted data with a real shared shift score under 0.9 for a held-out person, the instrument gets no claim.
- A leakage claim requires the random forest's random-split AUC to exceed its held-out-person AUC by at least 0.02.

The first run met the second part of the first rule: the planted shift scored 0.824 against a bar of 0.90. By its own rule that run was void, a verdict that says nothing for or against the claim. A calculation after the run showed the bar was out of reach: for the planted shift of 2 the best possible AUC is about 0.91 (section 5). A second preregistration, locked at 22:39 UTC the same evening, kept the real-data handling unchanged, tested planted shifts of 2, 3 and 4 on a fresh seed, and required the shift of 4 to reach 0.90 for a held-out person. If it passed, the first run's real-data results were to be credited. It passed. The second plan was written after the first run's results were seen, and its forecasts for the real data copy the first run's values; we report it as a repair, not as a blind test.

4 Data and provenance

The data are WESAD [1], obtained from the authors' page at the University of Siegen; the UCI repository record (doi:10.24432/C57K5T) points to that page. The terms allow scientific, non-commercial use with credit to the dataset's owners. This paper reports statistics only. Subjects are pseudonymous (S2 to S17; there is no S1 or S12); no attempt was made to identify anyone.

We used the chest device only, recorded at 700 Hz, read from the text files: ECG, EDA, EMG, skin temperature, three-axis acceleration and respiration. The pickled files were not opened. Condition times and the questionnaire answers came from each person's questionnaire file. The task was the binary one of the dataset paper: stress (the Trier test) against baseline and amusement. Windows of 60 s were cut every 15 s inside each condition. Windows with any missing feature were dropped, leaving 2,122 windows from 15 people, 30 percent of them stress.

Electrodermal activity (EDA), used throughout, is the electrical conductance of the skin, which rises with sweat gland activity under sympathetic control.

5 Methods

Features

Twenty features in six groups were computed per window. ECG: heart rate, the standard deviation of beat intervals, the root mean square of successive interval differences, and the ratio of the two as a rough balance measure. EDA: mean, standard deviation, slope and the count of skin conductance responses. Temperature: mean and slope. Respiration: breath count, amplitude and standard deviation. Acceleration: mean magnitude, its standard deviation and the mean of each axis. EMG: root mean square and 90th percentile of the rectified signal.

Models and score

Two classifiers were scored. A random forest averages 100 decision trees grown on resampled data with at least 3 windows per leaf. Logistic regression models the log odds of stress as a weighted sum of standardised features. The score is the AUC, the probability that a randomly chosen stress window gets a higher score than a randomly chosen non-stress window. With ranks r_i of the scores, n_1 stress windows and n_0 others,

$$A = \frac{R_1 - n_1(n_1 + 1)/2}{n_1 n_0} \tag{1}$$

where R_1 is the sum of the stress windows' ranks. AUC is reported two ways: global, over all held-out windows pooled, and within-person AUC, computed on each person's windows and then averaged over people.

Splits

- Random split over windows: five folds assigned at random to windows.
- Time blocks within a person: in each person's condition block, windows before the block's middle train and windows after it test, with a guard of 37.5 s on each side of the middle so that no test window overlaps a training window.
- Held-out person: train on 14 people, score the fifteenth, once for each person.
- Held-out person with each person's features scaled to that person's own baseline mean and standard deviation.

The leakage gap is the difference between the random-split and held-out-person global AUCs:

$$G = A_{rand} - A_{loso} \tag{2}$$

Planted data

The planted data keep the real windows' people, conditions and timing and replace the features. For window w of person s in condition block c, feature f is

$$x_{wf} = 0.3\varepsilon_{wf} + u_{scf} + d y_w \tag{3}$$

where ε_{wf} and u_{scf} are independent standard normal draws, the second shared by every window of one person's block, y_w is 1 for a stress window and 0 otherwise, and d is a planted shared shift added to the first feature only. With $d = 0$ each block has its own offsets in a random direction and nothing carries from one person to the next; a held-out person should score near 0.5. With $d > 0$ a held-out person has something real to find.

For one feature, the stress and non-stress values each have variance $1 + 0.3^2 = 1.09$, so the ceiling AUC for a shift is

$$C(d) = N\left(\frac{d}{\sqrt{2 \cdot 1.09}}\right) \tag{4}$$

where N is the standard normal cumulative distribution. At $d = 2$ this gives 0.912, which is why a bar of 0.90 left no room.

Sensor groups

Each sensor group, physiology without motion or breath (ECG, EDA and temperature) and motion alone (acceleration and EMG) were scored by held-out-person within-person AUC with the random forest. Whether each of the six single groups beats chance was tested with logistic regression against 100 block permutations, in which each person's stress label is moved to one of their condition blocks chosen at random, so block structure and the share of stress windows stay. The p value counts permutations at or above the real score, with one added to both counts:

$$p = \frac{k + 1}{n + 1} \tag{5}$$

The six p values were adjusted with the Benjamini-Hochberg correction, which keeps the expected share of false discoveries among the tests called significant at or below q 0.05. For p values sorted in increasing order, the ith of m is scaled to $mp_{(i)}/i$, and each adjusted value is the smallest scaled value at or after its place in the order.

Felt stress

For each person, the change in the six-item STAI total, a short state-anxiety questionnaire in which positive items are reverse scored, was taken from baseline to stress. The same person's change in mean EDA, heart rate and breath count from baseline windows to stress windows was compared with it across the 15 people by Spearman rho, the correlation of ranks.

Interval on the within-person mean

The second run reports a 95% interval on the mean within-person AUC by bootstrap by person: people are redrawn with replacement 2,000 times and the mean is recomputed each time.

6 Controls and planted checks

Planted data with nothing to learn ($d = 0$) were run in both runs. The random split read 1.000 (first run, 0.99999 before rounding) and 1.000 (second run). The held-out-person split read 0.469 and 0.391. Both passed their rules.

The planted shift failed in the first run, at 0.824 against the bar of 0.90 for $d = 2$. In the second run, on a fresh seed, the forest reached 0.692 at $d = 2$ (ceiling 0.912), 0.987 at $d = 3$ (ceiling 0.979) and 0.994 at $d = 4$ (ceiling 0.997). The rule on the shift of 4 passed. The value at $d = 3$ sits 0.008 above its computed ceiling; the ceiling counts one feature and two normal distributions, and we did not trace the difference. The two runs at $d = 2$ differ by 0.13, so this planted check is noisy at small shifts.

The block-shuffle control moves each person's stress label onto one of their own condition blocks at random and scores logistic regression on held-out people. Over 100 draws in the first run it gave a mean of 0.487 and a 95th percentile of 0.656 (p from draws 0.0099); over 40 draws in the second it gave 0.472 and 0.594 (p 0.024). The real logistic score of 0.896 clears both. This control's 95th percentile marks the top of its null distribution, the spread of scores the control gives over random draws.

7 Results

What the random split reports

A random split over windows scored an AUC of 1.000 and a held-out-person split scored 0.469. That is the planted data with nothing to learn, in the first run. The random forest separated stress from non-stress windows perfectly by recognising each block's offsets in windows it had already seen. Figure 1 sets these planted values beside the real ones.

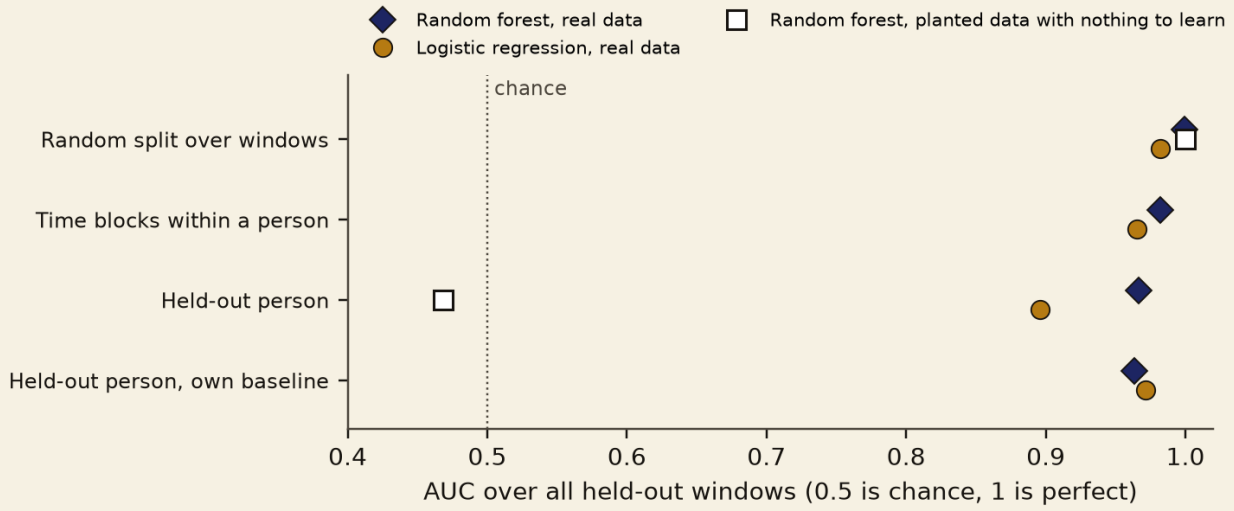


Figure 1. AUC over all held-out windows under each split. Diamonds: random forest on the real data. Circles: logistic regression. Open squares: the random forest on planted data with nothing to learn, which reads perfect under the random split and below chance for a held-out person. First run's result file.

The real recordings

Table 1 gives both models under the four splits. The random forest held across people at a held-out-person AUC of 0.966, against 1.000 under the random split. The second run, on a new seed, gave 0.999 and 0.966. The random split inflated the AUC by 0.033 for the random forest and 0.086 for logistic regression. Scaling each person to their own baseline brought the logistic score from 0.896 to 0.971 and left the forest at 0.964.

Table 1. AUC under each split, first run. Global AUC pools all held-out windows; within-person AUC is the mean over people of each person's own AUC.

Split	Forest, global	Logistic, global	Forest, within-person
Random over windows	1.000	0.982	1.000
Time blocks within a person	0.982	0.965	0.980
Held-out person	0.966	0.896	0.979
Held-out person, own baseline	0.964	0.971	0.979

The mean within-person AUC for a held-out person was 0.979, with a 95% interval of 0.961 to 0.994 by bootstrap over people. The worst person under the forest read 0.912 (S2); under logistic regression the worst read 0.646 (S15). Figure 2 shows every person and the planted-shift check.

Which sensors carry it

By sensor group, held-out person, random forest, the within-person AUCs were ECG 0.92, EDA 0.89, respiration

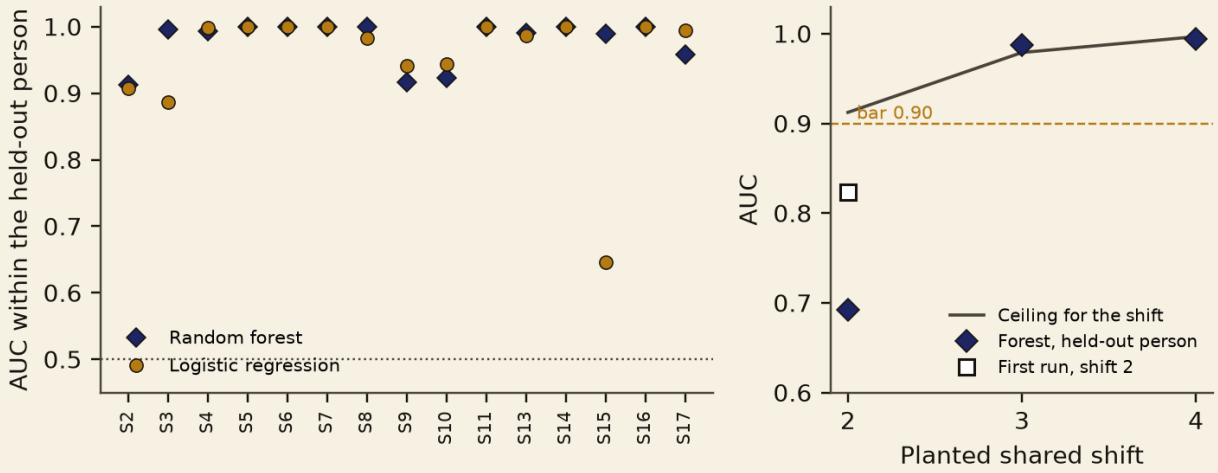


Figure 2. Left: AUC within each held-out person, both models. Right: the planted shared shift in the second run, held-out-person forest AUC against the ceiling of eq. 4, with the first run's failed value at a shift of 2 and the bar of 0.90.

0.85, acceleration 0.82, temperature 0.60 and EMG 0.45 (Figure 3). In the logistic permutation test five groups passed the correction at q 0.012 each; EMG did not (q 0.33). ECG, EDA and skin temperature together reached 0.97 for a held-out person, and acceleration with EMG reached 0.79. The Trier test has people stand and speak, so posture and breath move with the label. The classifier does not need them: physiology without motion or breath carries more than motion does.

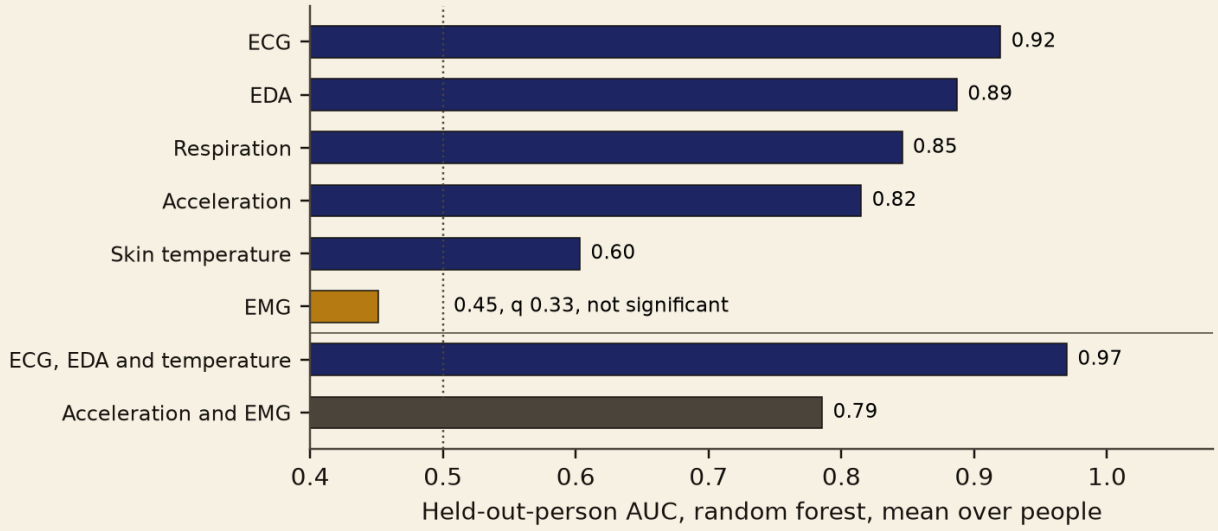


Figure 3. Held-out-person within-person AUC by sensor group, random forest. The lower two bars combine groups. The significance test used logistic regression and 100 block permutations; EMG's corrected q is printed beside its bar. First run's result file.

Felt stress against bodily change

The six-item STAI rose by 7.7 points on average from baseline to stress, so people reported feeling the test. Across the 15 people, the rise in reported anxiety did not follow the size of the bodily change: Spearman rho 0.03 with EDA, 0.02 with heart rate and -0.45 with breathing rate (p 0.09). The p values for EDA and heart rate were 0.93 each. With 15 people, a correlation's interval is wide, so this result rules out only a strong relation.

Forecasts against outcomes

Table 2 sets the forecasts in the first preregistration against the outcomes. The forecast that motion would carry more than physiology, given probability 0.6, was wrong.

Table 2. Forecasts fixed before any label was compared, against the first run. Forest values are global AUC except the sensor groups, which are within-person.

Quantity	Forecast	Outcome
Forest, random split	0.999	1.000
Forest, time blocks	0.97	0.982
Forest, held-out person	0.95	0.966
Forest, held-out person, own baseline	0.97	0.964
Logistic, held-out person	0.94	0.896
Random-split gap, forest	0.05	0.033
ECG, EDA, temperature, respiration	0.82, 0.80, 0.65, 0.85	0.92, 0.89, 0.60, 0.85
Acceleration, EMG	0.90, 0.75	0.82, 0.45
Physiology without motion; motion only	0.90; 0.92	0.97; 0.79
Planted, nothing to learn: random; held-out	0.95; 0.50	1.000; 0.469
Planted shift of 2, held-out person	0.98	0.824
Block-shuffle control mean	0.50	0.487
Felt-stress correlations	0.2, none significant	0.03, 0.02, -0.45; none significant
Groups passing correction	5 of 6	5 of 6

8 Discussion

The planted result is the general one. It needs no stress physiology: any dataset of overlapping windows with person or block structure will let a random split score the recording rather than the state. The argument between Saeb and colleagues [2] and Little and colleagues [3] was over which split estimates the error that matters. A known-answer test settles a narrower point for this design: when the planted answer is "nothing generalises", the random split says "perfect".

On the real data the inflation was small for the forest (0.033) and larger for logistic regression (0.086). Logistic regression gains the most from scaling each person to their own baseline, which suggests the gap for that model comes from differences in each person's resting level; we did not test that reading separately. The classifier does hold for new people inside this sample, which agrees with the dataset paper's held-out results [1].

The felt-stress result points the same way as the question Campbell and Ehlert [5] put: people reported more anxiety and their bodies changed, but across people the size of the reported rise did not track the size of the bodily change at any strength this sample could detect. A classifier trained on bodily signals reads the laboratory hour; it does not read how much that hour cost a given person.

9 Limitations

- Fifteen people, one laboratory, one protocol, about one hour each. A held-out person is a stranger only inside this sample.
- The stress session tends to follow baseline in time, so order is confounded with condition. A cardiac signal also carries posture and speech.
- With 15 people, the interval on any felt-stress correlation is about plus or minus 0.5. "Did not follow" means no relation of that size was detected.
- The first run's planted check failed through our own careless bar. The repair was written after seeing that failure, on a new seed, and its real-data forecasts copy the first run.
- The planted check at a shift of 2 gave 0.824 and 0.692 in the two runs, so it is noisy at small shifts.

- Nothing here reaches daily stress, chronic stress or trauma, and the classifier says nothing about what the hour meant to anyone.
- Features and models were fixed by us and not tuned; scoring was by our own code and was not blind.

10 Verified

Verified: on 9 November 2026 the 2,122 windows and 15 people, the stress share, every AUC in Tables 1 and 2, the two gaps, the within-person mean and its interval, each person's AUC, the per-sensor AUCs and their corrected q values, the planted values in both runs and their ceilings, both block-shuffle controls and the felt-stress correlations were re-read from the two runs' result files. They match the findings ledger, the dated append-only list of our results. Both preregistration SHA-256 values match those recorded in the result files, and the result files' leading hashes match the ledger. All three figures were drawn by a script that reads only the two result files. The study was not re-run for this paper. The cited papers were found by searches run on 9 November 2026 and read at the level of abstracts and search summaries, except Campbell and Ehlert, whose abstract was read during the study on 31 October 2026; Saeb and colleagues' paper was known through the two replies and the search summary, and its full text was not read.

11 Corrections

- 31 October 2026: the first run's planted check failed because its bar of 0.90 was above the ceiling of about 0.91 for a shift of 2. A second preregistered run repaired the plant. Both runs are reported.
- 31 October 2026: a review of our ledger recorded the entry as self-graded and not blind, with order confounded with condition. Both are stated in the limitations.
- 9 November 2026: first published version.

12 Reproduction

- Data: request WESAD from its authors under their non-commercial terms. Use the chest recordings and the questionnaire files.
- Cut the baseline, amusement and stress conditions into 60 s windows with a 15 s step and compute the twenty features of section 5. Drop windows with a missing feature.
- Score a random forest (100 trees, at least 3 per leaf) and a logistic regression on standardised features by AUC, eq. 1, under the four splits, and compute the gap, eq. 2.
- Build planted data with eq. 3 at shifts of 0, 2, 3 and 4 and check that a random split reads high and a held-out person near 0.5 at a shift of 0.
- Run the block-shuffle control and the sensor-group permutations, adjust with Benjamini-Hochberg, and compute the felt-stress Spearman correlations.
- First run: preregistration SHA-256 9263e28c14233905170468307af9706b5751c407f01bbb7d15ff253c4e6760c5. Plan SHA-256 98c0135c808b80cb0bb6bd7222184b1868739ac0f6d2cf1626b2b954c17484d6. Run 20261031T223719Z. Result SHA-256 65165c7a7ae8e10e608678e849959846e47f195c143fcb9db6665e22021d8296.
- Second run: preregistration SHA-256 5f1b6cda094184677b55ab08c84f05613e04f30f8c3a5484ec725a2b18b545c2. Plan SHA-256 5b45f5620a043beb623d96d6427e3b2e3024aeacef4ceb867f96e2164e9d6fa1. Run 20261031T223929Z. Result SHA-256 142aab473f4fbe40d2680a617c33962bf7ff717915b811b45f3b9cbdc5db6aae.
- Cost: the analysis and control took about 100 s in the first run and 31 s in the second on one desktop CPU, after feature extraction.

13 Appendix A. The preregistered verdicts

Each verdict is a set of rules written before the run, read off by code. A verdict passes only when every rule holds. Values are read from the two result files.

Table 3. Preregistered verdicts and the rule that decided each one.

Verdict	Deciding rule	Value	Holds
First run: planted, nothing to learn, random	random-split AUC at least 0.8	1.000	yes
First run: planted, nothing to learn, held out	held-out-person AUC at most 0.6	0.469	yes
First run: planted shift	shift of 2, held-out-person AUC at least 0.9	0.824	no
First run: leakage	forest gap at least 0.02	0.033	yes
First run: control	logistic held-out AUC above the block-shuffle 95th percentile (0.656)	0.896	yes
First run: physiology above motion	physiology-only AUC above motion-only AUC (0.786)	0.970	yes
Second run: planted shift	shift of 4, held-out-person AUC at least 0.9	0.994	yes
Second run: planted, nothing to learn, random	random-split AUC at least 0.8	1.000	yes
Second run: planted, nothing to learn, held out	held-out-person AUC at most 0.6	0.391	yes
Second run: leakage	forest gap at least 0.02	0.033	yes
Second run: control	logistic held-out AUC above the block-shuffle 95th percentile (0.594)	0.896	yes

14 Glossary

Each term is defined where it is first used in the text. The same definitions, with the field each comes from and what each excludes, are collected here.

- **Trier Social Stress Test**. A laboratory stressor in which a person prepares and gives a short public talk and then does mental arithmetic aloud in front of an evaluating panel. Field: psychology and psychoneuroendocrinology. Excludes daily stress, chronic stress and trauma.
- **random split over windows**. Assigning windows to training and test sets at random, so that windows from the same person and the same recording block fall on both sides. Field: machine learning evaluation; called record-wise cross-validation in the clinical literature. Excludes any split that keeps all of one person's data on one side.
- **overlapping windows**. Fixed-length stretches cut from a recording at a step shorter than their length, so that neighbouring windows share samples; here 60 s windows every 15 s share 75 percent of their samples. Field: signal processing. Excludes windows cut end to end with no shared samples.
- **preregistration**. A dated written plan for a study, fixed and hash-locked before any result is seen: the question, the data, the statistic, the controls, the forecasts and the kill conditions. Field: research method (metascience). Excludes a plan revised after looking at the outcome, and a protocol written only for the record after the run.
- **planted check**. Data built so that the right answer is known, run through the same analysis before the real data, to show that the analysis can see the effect when it is there. Field: research method; called a positive control in experimental science. Excludes any check made on the real data, which cannot show what the analysis misses.
- **control**. A version of the data with the structure under test removed (for example shuffled or pooled), scored the same way as the real data, so the real score has something to beat. Field: statistics and experimental

design. Excludes a comparison with a different dataset or with a published number.

- **kill condition**. A rule written in the preregistration that, when met, decides against the claim with no room for judgement after the run. Field: research method (our usage, after Popper's falsification). Excludes reasons found after the run for setting a result aside.
- **void**. The verdict of a run whose planted check failed or whose sample fell below its stated minimum; it says nothing for or against the claim. Field: research method (our usage). Excludes a null result, which comes from a working instrument.
- **electrodermal activity (EDA)**. The electrical conductance of the skin, which rises with sweat gland activity under sympathetic nervous control. Field: psychophysiology. Excludes a direct reading of emotion; temperature and movement also change it.
- **random forest**. A classifier that averages many decision trees, each grown on a resampled copy of the training data with a random subset of features tried at each split; here 100 trees with at least 3 windows per leaf. Field: machine learning. Excludes any model tuned on the test person.
- **logistic regression**. A classifier that models the log odds of the positive class as a weighted sum of the features; here on standardised features with an L2 (ridge) penalty at strength C = 1. Field: statistics. Excludes interactions between features, which it does not model unless they are added by hand.
- **AUC (area under the ROC curve)**. The probability that a classifier gives a randomly chosen positive case a higher score than a randomly chosen negative case; 0.5 is chance and 1 is perfect separation. Field: statistics and machine learning. Excludes accuracy at one threshold, and any statement about how well the scores are calibrated as probabilities.
- **within-person AUC**. The AUC computed on one held-out person's windows alone, then averaged over people; it asks whether the classifier ranks that person's stress windows above their other windows. Field: machine learning evaluation, our usage. Excludes the global AUC, which pools all held-out windows and also rewards separating people from each other.
- **held-out person (leave one subject out)**. Training on all people but one and scoring on the one left out, repeated so that each person is held out once. Field: machine learning evaluation; called subject-wise cross-validation. Excludes a test on people from another study, lab or protocol.
- **planted shared shift**. A fixed amount added to one feature in every stress window of every person in planted data, in the same direction for all, so that a classifier tested on a new person has something real to find. Field: research method, our usage. Excludes the per-person, per-block offsets, which point in a random direction and carry nothing from one person to the next.
- **ceiling AUC for a shift**. The AUC a perfect reader of the one shifted feature could reach when the stress and non-stress values of that feature are normal with a variance of 1.09 each (a block offset of variance 1 plus window noise of variance 0.09). Field: statistics, our derivation. Excludes the effect of the 19 features that carry no shift, and of the small number of people.
- **Benjamini-Hochberg correction**. A rule that adjusts a set of p values so that, among the tests called significant, the expected share of false discoveries stays at or below a chosen level q. Field: statistics, multiple testing. Excludes control of the chance of even one false discovery, which is a stricter rule.
- **STAI, six-item form**. A short form of the State-Trait Anxiety Inventory in which a person rates six statements about how they feel at that moment; positive items are reverse scored so that a higher total means more anxiety. Field: psychology. Excludes a clinical diagnosis, and anxiety as a lasting trait.
- **Spearman rho**. The correlation between the ranks of two quantities; it runs from -1 to 1 and measures how well one quantity's order follows the other's, whatever the shape of the relation. Field: statistics. Excludes the size of the change, and any causal direction.
- **95% interval**. The range that holds the middle 95 percent of the bootstrap or control draws for a number; we report it beside the number. Field: statistics. Excludes a statement that the true value lies inside with 95 percent probability.

- **bootstrap by person**. Redrawing people with replacement, 2,000 times, and recomputing the mean of the per-person scores each time; the middle 95 percent of those means is the interval. Field: statistics, resampling. Excludes redrawing single windows, which would treat one person's many windows as many independent people.
- **p value (from draws)**. The share of control draws scoring at or above the real score, with one added to the count and to the number of draws, so it is never zero. Field: statistics. Excludes a p value from a textbook formula that assumes a distribution.
- **null distribution**. The spread of scores a control gives over many random draws; its 95th percentile is the bar a real score must clear. Field: statistics. Excludes a theoretical distribution assumed rather than drawn from the data.
- **findings ledger**. The append-only, dated list of every result we have recorded, numbered F1, F2 and on; corrections are added as new dated lines and old entries are never rewritten. Field: our records. Excludes a curated list of successes.
- **verdict**. The pass or fail of one preregistered rule set against the run's numbers, read off by code. Field: research method (our usage). Excludes an interpretation written after reading the numbers.

15 References

[1] Schmidt, P., Reiss, A., Duerichen, R., Marberger, C. and Van Laerhoven, K. (2018). Introducing WESAD, a multimodal dataset for wearable stress and affect detection. Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI 2018). doi:10.1145/3242969.3242985. Data: UCI Machine Learning Repository, doi:10.24432/C57K5T; scientific non-commercial use with credit.

[2] Saeb, S. and others (2017). The need to approximate the use-case in clinical machine learning. GigaScience 6(5).

[3] Little, M. A. and others (2017). Using and understanding cross-validation strategies. Perspectives on Saeb et al. GigaScience 6(5). doi:10.1093/gigascience/gix020

[4] Chaibub Neto, E. and others (2019). Detecting the impact of subject characteristics on machine learning-based diagnostic applications. npj Digital Medicine. doi:10.1038/s41746-019-0178-x. Preprint arXiv:1712.03120.

[5] Campbell, J. and Ehlert, U. (2012). Acute psychosocial stress: does the emotional stress response correspond with physiological responses? Psychoneuroendocrinology. doi:10.1016/j.psyneuen.2011.12.010