

Beyond Mean Temperature: Measuring Cognitive Stress and Boredom by Facial Infrared Thermography

Shuja Rehman

`shuja_r@ar.iitr.ac.in`

Indian Institute of Technology Roorkee

Prabhjot Singh Chani

Indian Institute of Technology Roorkee

Rajasekar Elangovan

Indian Institute of Technology Roorkee

Article

Keywords: Infrared thermography, Measurement uncertainty, Contactless measurement, Autonomic nervous system, Classification, Leave-one-subject-out validation

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10940616/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Facial infrared thermography (IRT) is a contactless index of autonomic state, but its ability to separate two engaged cognitive states is uncharacterised. This study asked whether facial IRT can discriminate cognitive stress from boredom, and which thermal features carry that distinction. Eleven participants completed a Montreal Imaging Stress Task and a boredom-inducing memory task, while thermal images of 24 facial regions were acquired every 30 s. Six pixel statistics per region produced 144 instantaneous features and sixteen temporal summaries described their variation over time, all baseline-referenced within participant. Inference used exact within-participant sign-flip enumeration; classification used leave-one-subject-out validation with tuning confined to training folds. Absolute temperature did not discriminate the tasks (omnibus $P = 0.47$), and its 0.191°C between-task contrast sat below the instrument's 0.3°C black-body uncertainty. The frame-to-frame rate of change separated the tasks in 142 of 144 region-by-statistic series and in all 11 participants ($d_z = 1.19$, exact $P = 0.001$). A two-feature linear classifier identified 20 of 22 sessions and ranked all 11 participants correctly; a 14-model by 6-feature-set sweep showed that feature choice governed performance. Thermal dynamics, not mean temperature, provide the resolvable signal for contactless cognitive-state assessment.

Background

Sustained cognitive states such as stress, mental fatigue and boredom influence performance, safety and wellbeing in workplaces, classrooms and vehicles, yet they are still assessed largely through questionnaires that interrupt the activity under study. Facial infrared thermography (IRT) offers a contactless alternative and is increasingly used as an index of the peripheral vasomotor state and, through it, of autonomic activity ^{1,2}. The underlying rationale is direct: sympathetic outflow to apical facial skin closes arteriovenous anastomoses (AVAs), reduces local blood flow and lowers surface temperature, so a calibrated thermal camera can track autonomic activity without touching the participant ^{3,4,10}. A substantial methodological literature has developed around this principle, providing reporting guidelines, uncertainty estimates and standardisation protocols for the instrument itself ^{5,6,7}, together with tracking pipelines that recover facial temperature changes within a few seconds under free head movement ^{8,9}.

One question, however, has received comparatively little attention. Facial IRT is most often used to detect whether an autonomic response is present at all, by comparing an aroused state against a resting baseline. In many applied settings the relevant question is different: whether the instrument can separate two active cognitive states from each other. In operator monitoring, educational technology, indoor environmental assessment and clinical psychophysiology, the reference condition is rarely rest but sustained cognitive engagement, and the two states of interest may sit close together in autonomic space. Because the uncertainty of the instrument places a lower bound on what can be resolved, and because that bound differs across the families of features derived from a single recording, the choice of feature family may determine whether such a discrimination succeeds at all.

Sympathetic control of facial skin temperature

From psychological stress to facial vasoconstriction

The route from psychological stress to facial temperature change has been characterised at both circuit and peripheral levels ⁴. Psychological stress engages a sympathetic pathway originating in the thoracolumbar spinal cord ¹¹. Rodent work has identified a circuit specific to psychological stress, running from the dorsal peduncular cortex through the dorsomedial hypothalamus to the medullary raphe; selective pharmacological inactivation of it blocks stress-induced temperature change without affecting other thermoregulatory or cardiovascular responses ¹². The facial thermal response to stress is therefore specific rather than a by-product of general arousal.

At the effector level, noradrenaline is released at terminals surrounding arteriovenous anastomoses in apical skin. These thick-walled shunts allow blood to bypass the capillary bed, and their closure under sympathetic activation sharply reduces local blood flow and lowers surface temperature ^{3,10,13}. Two timing constraints follow. Microcirculatory vasoconstriction must persist for at least five seconds before a surface temperature change becomes detectable, and in primate work a significant nasal temperature change emerges within 10 to 30 seconds of stimulus onset ^{14,15}. Facial thermal responses are therefore slow relative to electrodermal or cardiac measures, which constrains the temporal resolution available to any IRT-based assessment.

Vascular territories and region selection

Apical facial skin contains a dense network of AVAs under noradrenergic control and is consequently highly responsive to sympathetic outflow ^{3,4}. The dual arterial supply of the face shapes the spatial distribution of thermal responses: the nose, cheeks, lips and temples are supplied by the external carotid artery and are preferentially affected by sympathetic vasoconstriction ¹⁶, whereas the periorbital and forehead regions are supplied by the ophthalmic branch of the internal carotid and are relatively protected. The 24 regions used here sample both territories under identical conditions.

Physiological basis of the thermal feature families

Sympathetic outflow to apical skin arrives in irregular bursts directed at spatially discrete groups of AVAs, and blood flow in AVA-dense regions varies by a factor of five to ten between its maximum and minimum (¹⁷, cited in ⁴). Because these bursts arrive asynchronously, adjacent AVAs within a region do not close simultaneously, and two observable consequences follow. In the spatial domain, the thermal field within a region becomes non-uniform, recordable as an increase in the standard deviation of pixel temperatures. In the temporal domain, the regional value fluctuates between frames, recordable as an increase in the frame-to-frame rate of change of the regional mean. Neither quantity is a temperature

difference, and neither is bounded by the sensor's temperature-difference uncertainty in the way an absolute level is ⁶.

Three families of thermal features

Any facial IRT recording can be summarised through three largely independent families of features. The absolute-temperature family comprises the mean, median, minimum and maximum of pixel temperatures within each region, and inherits the sensor's temperature-difference uncertainty in full. The spatial-heterogeneity family comprises the standard deviation and range of pixel temperatures within a region; the temporal-variance family comprises the frame-to-frame rate of change of any regional statistic. Both are derivatives, one spatial and one temporal, and both are invariant to a constant offset. The three families are therefore bounded by different components of the instrument's uncertainty budget, and the physiology outlined above predicts that they differ in what they can resolve. This study characterises what each resolves for the specific case of discriminating cognitive stress from boredom.

Interpretable machine learning provides the link between the thermal features and the cognitive-state decision. Interpretability matters here for a metrological reason as well as a modelling one, since the decision rule of an instrument should be traceable from the sensor reading, through the derived quantity, to the output. A model whose coefficients correspond to named physical quantities is the simplest object that satisfies this requirement, and it is the approach adopted here. Reported classification accuracies in thermal affective computing span roughly 48 to 99 per cent ^{1,18,19}, a spread that is to be expected when features and classifiers are selected without domain constraint. Constraining both in advance, on physiological grounds, should narrow that variability.

Baseline correction and subject-level validation are prerequisites for any small-sample study of this kind. Between-participant standard deviations in resting baseline temperature reach 0.6 to 2.0°C across facial regions ²⁰, so per-participant standardisation is essential. Leave-one-subject-out validation is similarly necessary for small thermal samples, because frame-level or record-level splitting inflates performance estimates by allowing individual-specific thermal patterns to appear in both training and test data ²¹. Both procedures are applied throughout.

Cognitive stress and boredom as a discrimination problem

Most thermal affective research has addressed categorical basic emotions, or the detection of acute stress against a resting baseline. Comparatively little work has attempted to separate two sustained states that both involve extended sedentary engagement with a screen, although that contrast is the ecologically relevant one: in operator monitoring, educational technology, indoor environmental assessment and clinical psychophysiology, the comparison is rarely between an extreme state and rest, but between graded states arising during sustained cognitive engagement ^{22,23,24,25}. Boredom in particular is best characterised not as an absence of arousal but as a specific under-aroused and

restless state, associated with attentional lapses and heightened mind-wandering²³. Thermal imaging has proved sensitive to comparably subtle affective processes, such as guilt in children detectable in the absence of behavioural markers²⁶. The few thermal studies of low-engagement states such as mental workload or fatigue have done so in isolation, without a matched high-demand comparison in the same participants^{27,28}.

The comparison tested in this study is between a mental arithmetic stressor and a boredom-inducing visual memory task. As reported in Results, the manipulation check confirms that the memory task was perceived as monotonous and that it differed sharply from the stressor in reported cognitive demand, although it did not produce reliable self-reported boredom in every participant.

Research gaps and objectives

Three gaps follow from the literature reviewed above. First, the resolution of facial IRT has been characterised against resting baselines but not for the discrimination of two engaged cognitive states, which is the contrast that applied settings actually require. Second, the sensor uncertainty of the instrument is well documented, yet its consequences for the choice among absolute, spatial and temporal feature families have not been examined empirically. Third, reported accuracies in this field vary widely, in part because features and classifiers are commonly selected after inspecting the data, and the cost of that selection is seldom quantified.

This study addresses these gaps with four objectives: (1) to determine which family of facial IRT features discriminates cognitive stress from boredom in a within-subject design, and which does not; (2) to compare the recovered between-task contrast against the black-body uncertainty of the instrument⁶, so that the outcome can be interpreted in metrological terms; (3) to quantify every analytic degree of freedom, including the multiplicity cost of screening 144 features, the selection cost of choosing a classifier after seeing the data, and the search cost of a sweep across 14 model families and 6 feature sets; and (4) to deliver a compact interpretable pipeline in which the decision reduces to two named physiological quantities and a linear Shapley attribution, so that the path from sensor to decision remains auditable.

Results

Manipulation checks

Self-report confirmed that the two tasks induced sharply different states. On the NASA Task Load Index, four of the five dimensions separated the conditions after Holm correction within the instrument (Table 1). The stress task was rated markedly more mentally demanding (mean difference 2.64, $d_z = 1.94$, all 11 participants in the same direction, Holm-adjusted $P = 0.005$) and more hurried and time-pressured (2.45, $d_z = 2.18$, 11 of 11, $P = 0.005$). It was also rated more effortful (2.27, $d_z = 1.13$, $P = 0.031$) and more frustrating on the combined item covering feeling insecure, discouraged, irritated, stressed and

annoyed (2.36 , $d_z = 1.20$, 9 of 11, $P = 0.023$). Perceived success did not differ ($d_z = 0.07$, $P = 0.902$), the expected result given that the arithmetic task adapts its difficulty to hold the success rate constant.

Task-specific items further corroborated the stress induction directly. During the stress task participants reported high felt stress (mean 4.27 of 5 , one-sample test against the scale midpoint $P = 0.002$) and high time pressure (4.55 of 5 , $P = 0.004$). General-state ratings ran in the same direction, with higher reported stress after the stress task (mean difference 0.82 , $d_z = 0.70$, exact $P = 0.063$).

The boredom induction, however, was milder, and the pattern bounds the interpretation of every thermal result that follows. The task was recognised as monotonous, with a mean rating of 3.55 of 5 for being repetitive or monotonous. Felt boredom, however, averaged exactly the scale midpoint at 3.00 of 5 and did not differ from it ($P = 1.000$); the eleven ratings were bimodal, with four participants rating their boredom at 4 or 5 and five at 1 or 2 . Overall, the task reliably produced low cognitive demand and was perceived as monotonous, but it produced felt boredom in only a subset of participants. The contrast tested thermally is a contrast in cognitive demand, and it is described that way throughout.

Circumplex ratings were available for a subset of the participants only. They placed the stress task at neutral arousal with negative valence (arousal 0.00 , valence minus 0.69) and the boredom task at low arousal with neutral valence (arousal minus 0.63 , valence 0.00). Neither axis reached significance at this reduced sample size (arousal $P = 0.156$, valence $P = 0.109$). The design does not, then, establish a pure manipulation of arousal, and any interpretation in terms of sympathetic activation is compatible with these data rather than demonstrated by them. Comfort with the room lighting was identical across conditions, with a mean difference of exactly zero and nine of eleven participants giving the same rating in both conditions, indicating that the manipulation acted on affective state rather than on the perceived environment.

Table 1

Subjective manipulation checks. Paired differences are stress condition minus boredom condition. Holm correction is applied within each instrument. Exact P values come from enumeration of all within-participant sign flips.

Instrument and item	Mean diff.	dz	Same direction	Exact P	Holm P
NASA-TLX: mental demand	+ 2.64	1.94	11 of 11	0.001	0.005
NASA-TLX: hurried or rushed pace	+ 2.45	2.18	11 of 11	0.001	0.005
NASA-TLX: insecure, irritated, stressed	+ 2.36	1.20	9 of 11	0.008	0.023
NASA-TLX: effort required	+ 2.27	1.13	7 of 11	0.016	0.031
NASA-TLX: success achieved	-0.18	0.07	4 of 11	0.902	0.902
General state: stressed now	+ 0.82	0.70	5 of 11	0.063	0.188
General state: alert now	+ 0.45	0.44	4 of 11	0.313	0.625
General state: lighting comfort	0.00	0.00	1 of 11	1.000	1.000
Circumplex: arousal	+ 0.63	0.65	5 of 8	0.156	<i>n/a</i>
Circumplex: valence	-0.69	0.86	1 of 8	0.109	<i>n/a</i>
<i>Note. NASA-TLX is scored - 3 to + 3, not the standard 0 to 100. "Same direction" counts strict non-ties in the group direction. Single-condition items are reported in the text.</i>					

Baseline characteristics and data quality

Resting baseline facial temperature averaged 34.87°C across the 24 regions, within the 33 to 37 degree range established for facial thermography²⁰. Baseline records comprised 5 to 21 frames per participant with a median of 13. Every recorded task block extended past minute 15, so the full 15-minute block structure was available in all 22 sessions and no session required exclusion or imputation.

Condition-independent thermal shift

Relative to the reading baseline, mean facial temperature rose by 1.22°C during the task blocks. The rise was present in all eleven participants, with a range across participants of 0.26 to 2.37°C (exact P = 0.001). It was of the same magnitude in both conditions and thus carries no information about which task was being performed. Its origin is not identified by this design, because the reading baseline always preceded both task blocks, which confounds acclimatisation, task onset and postural change with elapsed time.

The magnitude of this shared shift also matters for what follows. Expressed in matched units, any comparison based on absolute temperature must resolve a between-condition difference against a common movement 6.4 times larger than itself, where the between-condition difference is the mean absolute regional contrast of 0.191°C. For the rate-of-change measure the equivalent ratio is 1.2,

because that statistic is invariant to a constant offset. This asymmetry is quantified below in the comparison of dynamics and level, and discussed under Alternative explanations.

Instantaneous features

All 144 instantaneous features were screened for a condition difference in session means (Fig. 3a). Six reached $P < 0.05$ and all six are spatial heterogeneity measures; none of the 96 absolute-temperature features reached significance in any region (Fig. 3b, Table 2). This exclusive concentration of nominal hits in the heterogeneity statistics is the pattern predicted in the Introduction, and it is improbable under a null in which the two families are equally sensitive, since heterogeneity statistics make up only one third of the feature set.

Three cautions accompany Table 2. First, nothing survives correction for multiplicity, with a minimum Benjamini-Hochberg adjusted P of 0.281, exactly as the arithmetic set out in Methods predicts. Second, the six entries are three regions crossed with two closely related statistics, not six independent findings; the matrix of eleven participants by 144 contrasts places 72.1 per cent of its variance on a single component, with a participation ratio of 1.88, so these 144 measurements span approximately one effective dimension. Third, the six divide by side rather than by vascular territory: all four left-side features fall under stress and both right-side features rise, and the upper right cheek rises although it lies in the same external carotid territory as the upper left cheek, which falls. Restating the analysis on bilateral averages leaves no instantaneous feature significant. These six are therefore reported as a description of where the nominal differences lie, and no anatomical interpretation is placed on them.

The apparent advantage of heterogeneity over absolute temperature nonetheless requires careful statement. The ratio of mean standardised effects is 1.88, with heterogeneity at 0.317 and absolute temperature at 0.168. Referred to its own exact sign-flip null, however, that ratio is not significant, since the null itself has a median ratio of 1.26 ($P = 0.24$). The defensible claim is the categorical one, that all nominal hits fall in the heterogeneity family and none in the absolute-temperature family, and not the quantitative one that heterogeneity carries roughly twice the effect.

The omnibus test likewise confirms the absence of any level effect. The sum of squared t statistics across all 144 session-mean features was 133.2, with an exact P of 0.469 (Fig. 3c). On the evidence of this sample, the instantaneous level of the facial thermal field does not distinguish these two conditions.

Table 2

The six instantaneous features reaching $P < 0.05$ in the exploratory screen of 144 session-mean contrasts.

Feature	dz	t (10 df)	Exact P	Side	Direction under stress
Lower left cheek, standard deviation	-1.03	-3.42	0.002	Left	Lower
Lower left cheek, temperature range	-0.83	-2.76	0.020	Left	Lower
Right eye hood, standard deviation	+ 0.77	+ 2.55	0.028	Right	Higher
Upper left cheek, temperature range	-0.78	-2.57	0.028	Left	Lower
Upper left cheek, standard deviation	-0.69	-2.27	0.033	Left	Lower
Upper right cheek, standard deviation	+ 0.68	+ 2.26	0.041	Right	Higher
<i>Note. All six are heterogeneity statistics; none of the 96 absolute-temperature features reached $P < 0.05$. No feature survives Benjamini-Hochberg correction (minimum adjusted $P = 0.281$).</i>					

Temporal features

The temporal domain, in contrast, yielded a considerably stronger result. The frame-to-frame rate of change separated the two conditions in 142 of the 144 region-by-statistic series and in every one of the eleven participants (mean difference 0.183, $d_z = 1.19$, $t(10) = 3.96$, exact $P = 0.001$), the smallest P value obtainable from a design of this size. Expressed in °C on the regional means, the thermal field moved by 0.432°C between consecutive frames during the stress task and by 0.280°C during the boredom task, so the difference is a matter of the boredom task holding the field markedly steadier rather than of stress alone driving it. The comparison of dynamics and level below shows that both figures sit below the reading baseline, so the two conditions are better described as differing in how far each suppresses thermal volatility relative to rest.

Because a single favourable analytic choice would not be convincing, Fig. 4 and Table 3 report the same test under nine further treatments of the series. On raw °C, with no normalisation, the effect holds in 144 of 144 series and 10 of 11 participants ($d_z = 1.06$, $P = 0.002$). After a linear trend is removed from every series it is essentially unchanged ($d_z = 1.16$, 11 of 11, $P = 0.001$). The result does not depend on a small number of extreme frames: replacing the mean absolute successive difference by its median gives $d_z = 1.50$, a 20 per cent trimmed mean gives $d_z = 1.53$, and discarding the single largest step in each series gives $d_z = 1.34$, all at P of 0.002 or below.

In addition, three treatments address the unequal frame spacing between conditions. Dividing each successive difference by the elapsed seconds between frames weakens the effect without removing it ($d_z = 0.83$, 9 of 11, $P = 0.020$). More directly, both conditions were resampled onto a single common regular time grid inside minutes 2 to 10, so that cadence and window are identical by construction; the effect holds on a 30-second grid ($d_z = 1.23$, 10 of 11, $P = 0.002$), a 34-second grid ($d_z = 0.89$, 9 of 11, $P = 0.015$) and a 40-second grid ($d_z = 0.94$, 9 of 11, $P = 0.013$). The effect is not, then, an artefact of the sampling asymmetry, but the unanimity across participants is: on a matched grid it is 9 or 10 of 11 rather than 11 of 11, and the honest summary of the effect size is a range from 0.83 to 1.23 rather than a single value.

No single participant drives the result. Repeating the primary directional test eleven times, dropping one participant each time, gives d_z between 1.13 and 1.33 with the direction unanimous among the remaining ten in every refit. Task order is answered by the same result without a separate test: because all eleven participants moved in the same direction, the effect necessarily holds within both order subgroups, in five of five participants who received the stress task first and six of six who received the boredom task first.

The omnibus test on the rate-of-change vector gives a sum of squared t statistics of 718.7 with an exact P of 0.004, against 133.2 and $P = 0.469$ for the corresponding level vector (Fig. 3c). At the level of individual features, 62 of 144 reached $P < 0.05$ and 12 survived Benjamini-Hochberg correction, with a minimum adjusted P of 0.038; aggregated to whole regions, 12 of 24 reached $P < 0.05$ and 8 survived correction. Two qualifications belong with this. Ten of the twelve surviving features are absolute-temperature statistics rather than heterogeneity statistics, so the temporal effect is not confined to the heterogeneity family in the way the instantaneous effect is. And the surviving features concentrate in the periocular regions and temples, which are the regions most exposed to viewing-angle change. Restricting the screen to the ten mid-face regions removes every survivor: 20 of 60 features reach $P < 0.05$ but none survives correction, and the minimum adjusted P rises to 0.059. The twelve-survivor result is consequently not carried forward as the study's principal feature-wise evidence. The aggregate effect, by contrast, is if anything cleaner in the mid-face: restricted to those ten regions the global rate of change gives $d_z = 0.98$ in 10 of 11 participants ($P = 0.002$) with all 60 of 60 series directional, against 142 of 144 for the full set.

Table 3

Global rate of change, stress condition minus boredom condition, under ten treatments of the underlying series.

Treatment of the series	Mean diff.	dz	Participants	Series	Exact P
Baseline normalised (primary)	+ 0.183	1.19	11 of 11	142 of 144	0.001
Raw °C	+ 0.141	1.06	10 of 11	144 of 144	0.002
Linearly detrended	+ 0.184	1.16	11 of 11	142 of 144	0.001
Median instead of mean	<i>n/a</i>	1.50	10 of 11	<i>n/a</i>	0.002
20 per cent trimmed mean	<i>n/a</i>	1.53	11 of 11	<i>n/a</i>	0.001
Largest step discarded	<i>n/a</i>	1.34	11 of 11	<i>n/a</i>	0.001
Divided by elapsed seconds	+ 0.008	0.83	9 of 11	139 of 144	0.020
Common 30 s time grid	+ 0.196	1.23	10 of 11	<i>n/a</i>	0.002
Common 34 s time grid	+ 0.158	0.89	9 of 11	<i>n/a</i>	0.015
Common 40 s time grid	+ 0.183	0.94	9 of 11	<i>n/a</i>	0.013
Leave-one-participant-out (range)	<i>n/a</i>	1.13 to 1.33	10 of 10 in all 11 refits	<i>n/a</i>	0.002
<i>Note. Rate of change is the mean absolute frame-to-frame difference, averaged over all 144 series. The exact P-value floor for this design is 0.001. Last four rows address the frame-spacing asymmetry between conditions.</i>					

Comparison of dynamics and level on a common scale

A standardised effect can be large either because the underlying difference is large or because it is measured with little variation between people, and the distinction matters for how the main result should be described. Comparing the two families requires putting them in the same units, so the comparison was restricted to the 24 regional mean temperatures, which places both quantities in °C (Fig. 5, Table 4).

The between-condition contrast in absolute temperature straddles zero: it was positive in 6 of 11 participants ($d_z = 0.17$, $P = 0.578$), and the stress-to-boredom ratio of 0.995 exceeded one in the same 6 participants (Fig. 5a). Thermal volatility, in contrast, behaved differently. The between-condition difference in volatility amounted to 27.7 per cent of the baseline volatility of 0.546°C per frame interval ($d_z = 0.98$, 10 of 11, $P = 0.002$), and the ratio of stress to boredom volatility exceeded one in every participant (Fig. 5b), with individual ratios from 1.01 to 4.19.

Two features of that comparison require explicit statement. First, the ratio of 1.69 commonly quoted for this effect is the mean of 24 per-region ratios, which Jensen's inequality biases upward; the ratio of the corresponding means is 1.59, and the aggregate ratio of grand means is 1.55. The count of 11 of 11 belongs to the mean-of-ratios version only, and one participant sits at 1.01, indistinguishable from unity. The difference-based version of the same quantity is positive in 10 of 11. Second, and more important, volatility during both tasks was lower than during the reading baseline (Fig. 5c). Baseline volatility was 0.546°C per frame interval, against 0.432 during stress and 0.280 during the boredom task. The stress effect is thus a smaller reduction relative to rest rather than an increase above it. The most likely explanation is that the reading baseline is the first recording of the session and still carries residual thermal settling, which inflates its successive differences; the short and unequal baseline records are consistent with that. The between-condition contrast is unaffected, since it is a paired comparison of two task blocks, but no claim is made here that volatility tracks arousal in absolute terms.

In addition, a reliability advantage is present. Within each family, the ratio of the between-condition contrast to its between-participant standard deviation is 0.73 for volatility and 0.18 for level. This is a genuine measurement property and part of the explanation for the difference in standardised effect, and the Discussion sets out how far it goes.

Table 4

The two feature families compared on footings free of scale. Computed on the 24 regional mean temperatures only.

Comparison free of scale	Absolute temperature	Thermal volatility
Absolute value, reading baseline	34.87°C	0.546°C per frame
Absolute value, boredom task	<i>n/a</i>	0.280°C per frame
Absolute value, stress task	<i>n/a</i>	0.432°C per frame
Between-condition difference as a proportion of baseline	−0.49 per cent	+ 27.7 per cent
Standardised effect of that difference	$d_z = 0.17$	$d_z = 0.98$
Participants in the expected direction	6 of 11	10 of 11
Exact <i>P</i> value	0.578	0.002
Ratio, stress to boredom (mean of per-region ratios)	0.995	1.688
Ratio, stress to boredom (ratio of means)	0.995	1.547
Participants with mean-of-ratios above 1	6 of 11	11 of 11
Contrast divided by between-participant SD	0.184	0.731
Shared shift as a multiple of the contrast	6.4 times	1.2 times
<i>Note. The mean of per-region ratios is Jensen-biased upward; the ratio of means is the honest aggregate. The final row shows the shared shift each family must resolve its effect against.</i>		

Time course across the experimental block

The two-minute training period that opens each block is identical in structure across conditions and consequently functions as a negative control built into the protocol. For the rate of change the separation is clean: across the training period the difference between conditions is absent ($d_z = 0.01$, 4 of 11 participants positive, $P = 0.981$), while across the task window it stands at the design floor ($d_z = 1.19$, 11 of 11, $P = 0.001$). Re-cutting the recordings by elapsed time rather than by frame index confirms the same boundary from the other side: a planned contrast between the task period, minutes 2 to 10, and the questionnaire period that follows it, minutes 10 to 15, gives $d_z = 1.02$, exact $P = 0.008$.

Figure 6 shows the full time course in sliding four-minute windows. The difference is absent at the start of the block, rises through the task to a maximum near its middle, and has returned to zero by the time the questionnaires begin. It is a rise and decay confined to the task rather than a step function, and the decline begins before the task ends. No process running steadily through the session, such as instrument drift, chamber temperature, circadian warming or simple time on task, can produce that pattern.

The heterogeneity features, however, do not share this property. Tested individually in the training period, the lower left cheek standard deviation is already moving in the same direction as it does during the task, with only one of eleven participants running counter to it ($d_z = 0.64$, $P = 0.060$), and the right eye hood standard deviation is significant in the training period outright ($d_z = 0.49$, 10 of 11, $P = 0.040$). Table 5 gives the full comparison. A condition difference inside a period that is nominally identical across conditions is a session-level signal rather than a task-level one; plausible sources may include anticipation of the task about to begin, posture at the start of the block, and the position of the block within the visit. Task specificity is therefore claimed here for the rate-of-change measure alone.

Table 5
Each headline feature tested separately in the shared two-minute training period and in the eight-minute task window.

Feature	Training, d_z	Training, P	Task, d_z	Task, P
Global rate of change	+ 0.01	0.981	+ 1.19	0.001
Lower left cheek, standard deviation	-0.64	0.060	-1.03	0.002
Lower left cheek, temperature range	-0.54	0.103	-0.83	0.020
Right eye hood, standard deviation	+ 0.49	0.040	+ 0.77	0.028
Upper left cheek, standard deviation	-0.37	0.253	-0.69	0.033
Upper left cheek, temperature range	-0.36	0.257	-0.78	0.028
Upper right cheek, standard deviation	+ 0.12	0.690	+ 0.68	0.041
<i>Note. Only the rate of change is silent during the training period; task specificity is therefore claimed for the dynamic measure alone.</i>				

Classification

The fixed two-feature model classified 20 of 22 sessions correctly, which is 90.9 per cent with a Wilson 95 per cent confidence interval from 72.2 to 97.5 per cent (Fig. 7a). Recall was symmetric at 10 of 11 in each condition, and the two errors were one stress session and one boredom session belonging to different participants. Re-running the complete fitting procedure under all 2,048 within-participant label flips gives an exact P of 0.003. On the forced-choice measure, which compares each participant's two sessions against each other and so respects the clustered structure of the data, the stress session received the higher predicted probability in 11 of 11 participants (exact $P = 0.001$). A jackknife that dropped each participant in turn and refitted the whole pipeline returned between 16 and 19 of 20 correct, with the forced-choice ranking correct in all eleven repetitions. Four sessions sit within 0.09 of the decision boundary, so the absolute accuracy is more threshold-sensitive than the ranking is; the forced-choice figure is the more stable summary.

The fully automatic pipeline, by contrast, selects its own feature count and regularisation inside each training fold and uses no knowledge derived from the complete dataset. It reached 18 of 22, which is

81.8 per cent with an interval from 61.5 to 92.7 per cent, and 10 of 11 on forced choice ($P = 0.006$). The gap of two sessions between the two models is the cost of having identified the feature pair with knowledge of the full dataset, and it is priced in the Discussion.

Because the choice of classifier is itself an analytic degree of freedom, every combination of 14 model families and 6 feature sets was evaluated under the same nested procedure, giving 84 cells plus two reference points (Fig. 7b). Overall, the result is uniform in a way that matters for interpretation. Within the fixed two-feature set every family lands between 16 and 20 of 22, and seven families tie at the maximum of 20 with an identical per-participant error pattern, which indicates that they are all reading the same two numbers. No nonlinear family beats the best linear one anywhere in the grid: the best gradient boosting, random forest, support vector machine with a radial kernel and multi-layer perceptron results are all equal to or below the best logistic regression. Conversely, holding the family fixed and widening the feature set degrades performance sharply, from 20 of 22 on two features to 15 of 22 on all 144. The feature set decides the outcome and the model family does not, which is the expected behaviour when eleven participants supply the training signal^{21,41} and which is why no elaborate classifier is proposed here.

Three control analyses bound the interpretation of these figures (Table 6). A majority-class classifier scores 11 of 22 by construction. A model restricted to the 96 absolute-temperature features reaches 16 of 22 at best across five regularisation settings, and one restricted to the 48 heterogeneity features reaches 14 of 22. Most informative of all, a simple threshold on the global rate of change alone, with no fitted feature weights, scores 16 of 22 and ranks all eleven participants correctly on forced choice. A single number computed from the thermal record, with nothing learned from the training data except a cut point, therefore separates every participant's two sessions in the correct order. Figure 7c places these controls side by side. The ordering is the substantive result: any model built on the temporal measure beats any model built on the instantaneous level, and the gap between the best level-only model and chance is smaller than the gap between the level-only model and the two-feature model. Table 7 reports the primary set of four tests with Holm correction applied across them.

Table 6

Classification results and control analyses. All models use leave-one-subject-out validation with scaling, selection and tuning confined to the training folds.

Model	Features	Accuracy	95 per cent CI	Forced choice	Exact P
Fixed two-feature logistic	2	20 of 22 (90.9%)	72.2 to 97.5	11 of 11	0.003
Fully automatic nested pipeline	selected in fold	18 of 22 (81.8%)	61.5 to 92.7	10 of 11	0.006
Threshold on rate of change alone	1	16 of 22 (72.7%)	51.8 to 86.8	11 of 11	0.001
Absolute temperature features only	96	16 of 22 (72.7%)	51.8 to 86.8	9 of 11	not significant
Heterogeneity features only	48	14 of 22 (63.6%)	43.0 to 80.3	n/a	not significant
Majority class baseline	0	11 of 22 (50.0%)	30.8 to 69.2	tie	n/a
Earlier analysis of this dataset	over 1,000	17 of 22 (77.3%)	56.6 to 89.9	not reported	not reported
<i>Note. Forced choice compares each participant's two sessions against each other and is the appropriate subject-level summary. Single-family rows report the best of five regularisation settings and are ceilings rather than point estimates.</i>					

Table 7

The primary set, corrected by the Holm procedure across the four tests.

Primary test	Statistic	Exact P	Holm P
Global omnibus on rate of change	sum of squared t = 718.7	0.004	0.008
Directional test on rate of change	dz = 1.19, 11 of 11	0.001	0.004
Phase specificity, task against questionnaires	dz = 1.02, t(10) = 3.39	0.008	0.008
Forced-choice classification	11 of 11 participants	0.001	0.004
<i>Note. The four rows are not four independent claims; see the text. The 144-feature exploratory screen is corrected separately by Benjamini-Hochberg and yields no survivor.</i>			

Regional and statistical contributions

For the two-feature model the Shapley value of a feature is exactly its coefficient multiplied by the centred feature value, so the coefficients are reported directly. The global rate of change enters positively, pointing towards the stress condition; the lower left cheek standard deviation enters

negatively, so lower within-region heterogeneity at that site is associated with the stress condition. Both directions match the corresponding paired contrasts in Tables 2 and 3.

A frame-level attribution across all 144 instantaneous features was also computed. In that decomposition the nose emerges as the highest-contributing region, reproducing the pattern reported in the earlier analysis of this dataset. It is not reported here as a feature ranking, and the reason is structural. That model is fitted to 352 frames, but those frames comprise 16 from each of 22 sessions belonging to only 11 individuals, so the effective sample size is 22 for 144 free parameters; individual coefficients are not identified and any ordering among them is unstable. The instability is demonstrable rather than hypothetical: repairing the two timestamp defects recorded in Methods reorders the pixel-statistic ranking across a three-way near-tie.

Because the applied question is which region an instrumented system should track, regions were ranked instead on the measure that carries the effect (Fig. 8a). On the rate of change the strongest regions are the right eye hood ($d_z = 1.43$), the left iris (1.33) and the two temples (1.26 and 1.24), of which nine regions survive Benjamini-Hochberg correction. The nose ranks seventeenth of twenty-four and does not reach significance ($d_z = 0.58$, $P = 0.090$). The instantaneous and temporal levels of the analysis point to different regions, and the region that leads the frame-level attribution is not the region that carries the discriminating signal.

The empirically strongest regions are also those most exposed to viewing-angle change, so five region subsets defined by vascular anatomy rather than by that ranking were evaluated instead (Fig. 8b). The full 24-region set gives $d_z = 1.19$; the ten external carotid regions give 1.07 with all 60 of 60 series directional; the ten mid-face regions give 0.98, also 60 of 60; and a five-region set comprising the nose and the four cheek quadrants gives 0.97 with 30 of 30 series directional. The nose alone gives 0.58 and does not reach significance. A compact mid-face set therefore recovers most of the effect available in the full set, whereas the single canonical region does not. Across the six pixel statistics the effect is similarly insensitive to which is used (Fig. 8c): temperature range gives $d_z = 1.26$, minimum 1.24, standard deviation 1.13, mean 1.04, median 0.93 and maximum 0.75, with four of the six unanimous across participants.

Discussion

Measurement resolution and instrument uncertainty

The central claim of this study is not that a 90.9 per cent classifier is the best a facial IRT system can achieve on this task; deep-learning pipelines on posed-expression corpora reach comparable or higher accuracies^{42,43}. The claim is instead a metrological one. The three families of features of facial IRT identified in the Introduction differ in what they can resolve when comparing two engaged cognitive states, and the difference is bounded from below by the instrument.

The sensor-level black-body uncertainty of facial IRT is estimated at approximately 0.3°C by the systematic review of Stanic and Gersak ⁶, and viewing geometry contributes more measurement error than camera distance does ⁷. The between-condition contrast in the absolute-temperature family measured here is 0.191°C, which sits below the sensor's temperature-difference uncertainty floor. No amount of downstream modelling can resolve a level difference smaller than the sensor's temperature-difference uncertainty, and published claims that a level-based facial IRT feature discriminates two engaged cognitive states should thus be treated with caution unless the level contrast exceeds this floor.

The temporal-variance and spatial-heterogeneity families are not temperature differences. Both are derivatives of the field, one temporal and one spatial, and both are invariant to a constant offset in that field. Their uncertainty inherits the sensor's temperature-difference uncertainty differently: because they operate on differences within a session and within a region, the per-participant, per-region baseline standardisation applied here removes any static component of the sensor error and leaves only its within-recording variability. The rate-of-change measure recovers a between-condition effect of $dz = 1.19$ in every participant per frame, with a range of 0.83 to 1.53 across ten treatments of the underlying series including two that resample both conditions onto a common time grid. This provides a Type A estimate of the standard deviation of the rate-of-change family of the order of 0.15 in baseline-standardised units, and it is by construction of the same order as the between-condition contrast.

The interpretability of the discriminant that maps the recovered feature family to the cognitive-state decision is a metrological requirement, not only a modelling preference. A pipeline whose discriminant is opaque cannot be audited against the instrument's uncertainty budget. The two-feature linear-in-features discriminant reported here satisfies this requirement at three levels. At the feature-set level, the input is two named physiological quantities, both derivatives of the thermal field with named physiological correlates. At the model-family level, the 14 by 6 sweep (Fig. 7b) confirms that widening the feature set does not improve the outcome and that no nonlinear family exceeds the linear one anywhere in the grid. At the attribution level, the Shapley additive explanation for a linear-in-features model is exactly the product of the fitted coefficient and the centred feature value ^{39,40}: the two coefficients are themselves the attribution, and a zero-parameter threshold on a single physiology-derived measurement primitive already recovers 16 of 22 sessions and orders every participant correctly.

Physiological interpretation of the temporal response

The principal finding of this study is that facial thermography separates cognitive stress from boredom through the temporal dynamics of the thermal field rather than through its absolute level. This is not a departure from the physiological account set out in the Introduction; it is that account's more direct prediction.

The mechanism is likely to be pulsatile, in the sense set out in the Introduction. A measurement that averages the thermal field over a region and over a session is by construction blind to that pulsatility. Two observable consequences survive the averaging: within a frame, asynchronous closure leaves the

field spatially non-uniform; across frames, the arrival and decay of successive bursts leaves the regional value unsettled.

The earlier analysis of this dataset established the spatial consequence. The present analysis establishes the temporal one and finds it the stronger of the two by every measure applied. The temporal effect holds in every participant per frame, where the instantaneous screen yields only 6 nominal hits in 144 features; its omnibus P is 0.004 against 0.469; and it survives restriction to the mid-face regions with every series directional, which the instantaneous effect does not. That ordering is itself informative. Consider a burst brief relative to the 30-second sampling interval. It contributes to the temporal variance of the series whenever it occurs, but to the spatial variance of a frame only if that frame happens to be captured while the burst is in progress. The temporal measure integrates evidence that the spatial measure merely samples. The same logic underlies the wavelet decompositions that recovered tonic and phasic sympathetic components from nasal temperature where a single change score did not ⁴⁴, and the fluctuation analyses that separated stress from rest during dexterous tasks ^{45,46}.

The account above is written from the stress side, and the data support an equally direct reading from the boredom side. Boredom is characterised by attentional disengagement and reduced arousal rather than by a distinctive autonomic signature of its own ^{22,23,24}, and a withdrawal of the pulsatile sympathetic traffic that unsettles the thermal field predicts exactly what the Results report: the steadiest thermal field of the three states recorded here occurs during the boredom task, at 0.280°C per frame interval against 0.432 during stress and 0.546 during silent reading. On that reading the discrimination is carried as much by the quiescence of the bored state as by the agitation of the stressed one, and the vigilance-decrement literature, which reports a comparable fall in task engagement over time ²⁵, predicts the same direction, as does the thermal workload literature in which sustained low-demand monitoring is associated with a drift in facial temperature rather than a distinct signature ^{27,28}. The present design cannot separate these two descriptions, because with no resting condition interleaved between the task blocks there is no anchor against which to say which state moved. Both are recorded here because the contrast is symmetric and the paper should not be read as a study of stress alone.

No absolute-temperature feature discriminated the conditions, and the omnibus test on that family was null. Given the prominence of nasal cooling in this literature, that result requires a direct explanation rather than an apology.

Two considerations make an unchanged regional mean compatible with substantial sympathetic activity. The first is spatial: a facial region a few centimetres across spans more than one vascular territory, and if adjacent patches of skin move in opposite directions the mean of the region is nearly unchanged while its internal dispersion is not. The second is the size of the comparison. Canonical nasal cooling results are obtained against a resting baseline ^{16,20}, whereas the present comparison is between two active cognitive tasks. The Results show that both tasks displaced facial temperature 1.22°C from the reading baseline, the same order as reported stress-against-rest effects. The absolute-temperature family is not,

then, insensitive in this dataset. It responds strongly to task engagement and simply does not separate the two engaged states.

A third consideration is instrumental. As set out above, the 0.191°C between-condition contrast in absolute temperature sits below the approximately 0.3°C black-body uncertainty floor that Stanic and Gersak⁶ estimate from a review of 315 facial thermography studies, whereas the contrast in the rate-of-change family is not a temperature difference at all and is not bounded by it in the same way. The absence of a level effect in this dataset is accordingly consistent with the instrument as well as with the physiology, and it implies that separating two engaged cognitive states by absolute regional temperature may be beyond the resolution of facial thermography as currently practised, irrespective of sample size.

This reading aligns the present result with the earlier literature rather than opposing it. Nasal cooling remains a marker of sympathetic activation against rest. What these data indicate is that it is not a marker of which of two engaged cognitive states a person is in, and that the temporal structure of the signal is what carries that finer discrimination. That reading is consistent with the observation that nasal cooling tracks arousal even for mildly arousing material yet is uncorrelated with electrodermal response⁴⁷, which implies that no single facial index maps cleanly onto sympathetic activation in general, and with the wide spread of accuracies reported across thermal classification studies using different region sets and paradigms^{18,19}.

Scope and limits of the temporal measure

The rate-of-change statistic must not be described as vasomotion, and the reason is arithmetic. At a nominal sampling interval of 30 seconds the sampling rate is 0.033 Hz and the Nyquist limit is 0.0167 Hz. The cardiac band from 0.4 to 2 Hz, the respiratory band from 0.15 to 0.4 Hz, the myogenic band from 0.06 to 0.15 Hz and the neurogenic band from 0.02 to 0.06 Hz all lie above that limit⁴⁸, and Mayer waves near 0.1 Hz are six times above it⁴⁹. Only the lower part of the endothelial interval is resolvable at all, and sixteen samples spanning about eight minutes give a frequency resolution near 0.002 Hz across roughly eight usable bins. No spectral estimate is defensible from these data and no named oscillator can be assigned to the effect.

The statistic should therefore be read as a broadband measure of total variation whose spectral composition is unknown. Five candidate sources remain, and the present design cannot fully separate them. The first is slow redistribution of facial blood flow below 0.017 Hz. The second is aliased contribution from faster vasomotor activity, which skin attenuates but does not remove, and the acquisition applied no anti-aliasing filter. The third is discrete sudomotor activity, which produces transients rather than oscillations. The fourth is ocular-surface dynamics: a thermal camera pointed at the iris or inner canthus measures the tear film, whose temperature falls between blinks and resets at each blink, and blink rate changes with cognitive load. That candidate deserves weight here because the features surviving correction concentrate in the periocular regions, and because restricting the analysis

to mid-face regions removes every one of them. The fifth is movement and misregistration, discussed below.

Several observations constrain that list. The effect is unchanged by detrending, which argues against slow monotone redistribution as the sole source. It survives median, trimmed and largest-step-removed variants of the statistic, which argues against a small number of registration excursions. It survives resampling onto a common time grid, which excludes the sampling asymmetry. And it survives restriction to the mid-face with every series directional, which weakens but does not eliminate the ocular account, since head movement affects the whole field.

The timing of onset and offset constrains the mechanism usefully. The difference is absent from the shared training period, rises through the task and has decayed by the questionnaire period. That pattern excludes several candidates outright: cortisol requires twenty to thirty minutes to peak⁵⁰ and cannot switch on and off on this timescale, and neither can redistribution of core body heat, acclimatisation to the chamber, or postural change. What the pattern does not do is isolate sympathetic vasomotor activity in particular, because any process beginning and ending with the task remains available, including task-related movement, jaw activity, altered respiration and gaze behaviour.

Alternative explanations

An account requiring no physiology at all must be stated, because part of the observed advantage follows from the construction of the features. Both conditions displaced the level family 1.22°C from the reading baseline. In matched units an absolute-temperature comparison must thus resolve its between-condition difference against a shared movement 6.4 times larger, while for the volatility measure the corresponding factor is 1.2. Within-region standard deviation is a spatial derivative and the successive-difference statistic is a temporal one, and both are unchanged by the addition of a constant. Features of that kind were guaranteed a better signal-to-noise ratio than level features irrespective of any biology, and the Results quantify the resulting reliability advantage as 0.73 against 0.18.

The second account, in contrast, is movement. Both the within-region standard deviation and the mean absolute successive difference are inflated when a region shifts across a steep thermal gradient between frames, and a time-pressured arithmetic task plausibly involves more head, jaw and gaze movement than a memory game. This study contains no motion log and no static reference target, so the account cannot be excluded by design; modular thermal face-tracking pipelines with rigid landmark stabilisation can deliver up to 75 per cent improvement in the signal-to-noise ratio of ROI-extracted temperatures⁸, and deep-learning-assisted landmark tracking has been demonstrated to recover thermal changes on the order of a few seconds under free head movement⁹, so the instrumentation to test this directly now exists. It can, however, be quantified from within the data. A surrogate for rigid motion was constructed as the frame-to-frame disruption of the between-region spatial pattern, measured as one minus the rank correlation of the 24 regional mean temperatures across consecutive frames. That surrogate is higher during stress than during the boredom task (0.171 against 0.095, $d_z = 1.09$, $P =$

0.004), correlates with the rate-of-change difference at $r = 0.58$, and accounts for roughly half of its point estimate when adjusted for (0.088 against 0.183 unadjusted).

Two things must be said about that surrogate, and they pull in opposite directions. It shows the same phase structure as the effect itself, being null across the shared training period ($d_z = 0.01$, $P = 0.979$) and significant during the task, which rules out posture, camera state and session order as its source but is equally consistent with task-locked movement. And it is confounded with the signal it is being used to test, because genuine asynchronous anastomosis closure would also disrupt the between-region spatial pattern. The surrogate is therefore an upper bound on the movement contribution and not an estimate of it. The honest position is that a substantial share of the observed effect could be movement, that the present design cannot resolve what share, and that the instrumentation which would resolve it is specified under Limitations and future work.

Classification performance and analytic robustness

The fixed model reached 90.9 per cent and the fully automatic pipeline 81.8 per cent, against 77.3 per cent previously reported for this dataset. The automatic figure differs from the earlier one by a single session and should not be read as an improvement in its own right; the mechanistic changes responsible are the fixed protocol-based analysis window, the removal of the degenerate entropy features and the use of a single drift-invariant quantity in place of a large set of correlated temporal summaries. Both figures sit within the range reported in this literature, which spans roughly 48 to 98 per cent¹, but direct comparison against that range is of limited value because much of it is obtained without control for subject-level leakage, and cross-validated accuracy carries very wide error bars at small sample sizes regardless of the split⁴¹. Thermal infrared imaging has also expanded into a broad set of screening and monitoring contexts in recent years, extending well beyond the paradigms used to develop its interpretive framework², so cross-study comparisons must remain sensitive to the paradigm as well as to the sample size.

Three features of the present evaluation are more informative than the headline percentage. Selection and tuning occurred strictly inside training folds, so the reported figure is subject-independent. The forced-choice result of 11 of 11 respects the clustered structure of 22 sessions drawn from 11 individuals, which the session-level interval does not. And a threshold on a single drift-invariant number, with no fitted feature weights, already ranks all eleven participants correctly, which suggests the discriminating information is accessible without elaborate modelling. The 84-cell sweep reported in Results makes the same point from the other direction: deep and ensemble architectures that reach 93 to 96 per cent on large posed-expression databases^{42,43} have no advantage here, because the limiting quantity at eleven participants is the number of independent individuals rather than the flexibility of the classifier²¹.

That last point, together with the region ranking reported in Results, yields a concrete specification for an instrumented system, and it differs from what the frame-level attribution would have suggested. Five

mid-face regions, the nose and the four cheek quadrants, recover an effect of $d_z = 0.97$ with every one of their thirty region-by-statistic series directional, against $d_z = 1.19$ for the full twenty-four. The nose alone recovers $d_z = 0.58$ and does not reach significance, so the canonical single region is not sufficient. The quantity to record is the frame-to-frame change rather than the level, and it need not be an exotic statistic: the plain regional mean gives $d_z = 1.04$ with all eleven participants in the same direction. Personalised baseline normalisation remains necessary, because between-participant baseline variation is of the order of 0.6 to 2.0°C ²⁰ and would otherwise swamp a contrast of roughly 0.2°C .

The implication for measurement practice is specific. If the informative content of a facial thermal recording lies in its variation rather than its level, several conventions in this literature work against detection. Averaging within large regions suppresses the spatial component, averaging across long windows suppresses the temporal component, and reporting a single change score from baseline discards both. The sampling interval, often chosen for convenience, determines how much of the variation is visible at all.

This paper reanalyses a dataset previously reported by the same authors, and the relationship should be explicit. Three conclusions are retained: spatial heterogeneity carries all the instantaneous condition effects and absolute temperature carries none; the rate of change is the dominant temporal discriminator; and the classifier separates the conditions under subject-level validation. Two are revised. The magnitude of the heterogeneity advantage over absolute temperature was overstated, since the ratio of mean standardised effects is not significant against its own null ($P = 0.24$), so only the categorical claim survives. And the identification of the nose as the leading region, which came from frame-level attribution, does not hold on the measure that carries the discrimination, where the nose ranks seventeenth of twenty-four. One conclusion is withdrawn: the claim of dual temporal dynamics, in which heterogeneity separates the conditions immediately while absolute temperature diverges progressively, is not supported once the analysis window is fixed on protocol grounds, and the entropy features that contributed to the earlier session-level result are degenerate at this number of samples.

The selection costs of the present analysis are priced here rather than left implicit. The rate-of-change measure was chosen from fourteen candidate temporal summaries computed on these same data; under a fourteen-fold correction the directional test still holds. The two-feature pair was identified with knowledge of the full dataset, and priced against the complete search of candidate pairs it carries $P = 0.032$, which is why the fully automatic pipeline that never sees the full-data ranking is reported alongside it at 18 of 22 rather than 20 of 22. Approximately two sessions of the headline figure are attributable to hindsight, and the automatic pipeline's 18 of 22 is the more honest single-number summary of what a subject-independent system built on these principles achieves. A sweep across model families would not improve matters: across 84 combinations of model family (listed in Methods) and feature set, all evaluated with the same nested procedure, the best result was the same 20 of 22, but priced against the sweep it carries a selection-corrected P of 0.014 rather than 0.003, and after that correction a result of 18 of 22 obtained by sweeping would not have been significant. Two fixed models are consequently reported rather than the best of many.

The primary set in Table 7 should be read as a specification for replication rather than as an out-of-sample confirmation.

Limitations and future work

The design has real limits, all bounded by measurement or by sample size rather than by physiology. The boredom induction was milder than the stress induction, felt boredom averaged exactly the scale midpoint and the eleven ratings were bimodal, so the contrast is described throughout as one of cognitive demand with a partial boredom induction, consistent with the finding that boredom is a heterogeneous under-aroused state whose subjective and physiological expressions vary substantially between individuals²³. No autonomic or endocrine reference channel was recorded, so the classifier separates two tasks whose reported demand differs sharply and not two independently verified physiological states; nor is the contrast a clean manipulation of arousal, since neither circumplex axis reached significance in the subset of participants for whom ratings were available. Movement and misregistration remain the largest measurement threat this design cannot exclude, and Alternative explanations above quantifies rather than dismisses it: a within-data motion surrogate is elevated during stress and accounts for roughly half the point estimate on adjustment, while being confounded with the signal itself. A static left to right thermal asymmetry consistent with the directional daylight in the chamber is removed by construction by the per-participant, per-region baseline normalisation, but the six nominal instantaneous features divide by side rather than by vascular territory and do not survive bilateral averaging, so no anatomical interpretation is placed on them; view angle is in any case a documented source of error in facial thermography that contributes more than camera distance does⁷. Frame timing differs between conditions and although resampling onto a common time grid preserves the effect, it reduces the unanimity across participants from eleven to nine or ten; no spectral interpretation is possible for the reasons given under Scope and limits of the temporal measure; baseline records are short and unequal; and a single camera position and controlled environment were used throughout, so generalisation to other cameras, geometries or thermal environments is untested.

Inference is bounded by the sample size. Nothing survives correction across the 144 instantaneous features and this is arithmetic rather than a property of the data, since at eleven participants Holm across 144 requires a standardised effect above 1.60 and the exact null floors at a two-sided P of 0.00098; the twelve rate-of-change features that do survive correction across the full region set do not survive restriction to the mid-face, so the feature-wise evidence is weaker than the aggregate. Accuracy uncertainty is understated by the session-level interval because the 22 sessions form eleven clusters of two that share normalising constants, so the forced-choice figures are the appropriate subject-level summaries. Eleven participants support an existence claim rather than a generalisation claim, and the general consequences of low-powered designs, namely inflated effect-size estimates and reduced reproducibility, apply here as they do across the neurosciences⁵². A within-subjects design of this size detects a standardised effect of approximately 0.94 with 80 per cent power, which the primary effects exceed; it cannot establish that a classifier transfers to new individuals. Reaching a lower confidence

bound above 0.70 at a true accuracy of 0.85 would require approximately fifteen participants, and an interval half-width of 0.05 roughly one hundred.

The decisive next study is not a larger version of this one but a better instrumented one, and four changes would resolve what this design leaves open. Recording at 1 Hz or faster would bring the myogenic and neurogenic bands within the Nyquist limit and permit a genuine spectral decomposition, converting the broadband total-variation measure used here into an interpretable frequency-resolved one; individual anastomosis constrictions last at least five seconds ⁴, so a 1 Hz acquisition would resolve them directly. Rigid landmark tracking, a static high-emissivity reference target in the field of view, a motion log and a blink log would allow the movement and ocular accounts to be tested rather than bounded; the specific predictions are that the difference should survive motion-compensated re-registration and the exclusion of frames with head displacement above 2 millimetres, that the reference target should show no condition difference beyond the noise-equivalent temperature difference of the sensor, and that the mid-face effect should be independent of blink rate. Synchronous electrodermal and cardiac recording would test whether the excess frame-to-frame variation is carried by the number of thermal transients time-locked to electrodermal events rather than by their amplitude, which is the specific prediction of the pulsatile account. And a task that produces its intended state reliably across participants, in a sample large enough to estimate between-person generalisation, would convert the existence claim established here into the deployment claim that applications require.

Conclusion

This study has characterised facial infrared thermography as an instrument for discriminating cognitive stress from boredom. Three feature families of the same recording were compared under identical statistical treatment. The absolute-temperature family yielded no discriminating feature and produced a between-condition contrast of 0.191°C, below the sensor's black-body uncertainty floor of approximately 0.3°C ⁶. The temporal-variance family, operationalised as the frame-to-frame rate of change of the thermal field, is not bounded by that floor in the same way and recovers the discrimination in 142 of 144 region-by-statistic series and, per frame, in every participant.

Under leave-one-subject-out validation an interpretable two-feature discriminant classified 20 of 22 sessions correctly and ranked every participant's two sessions in the correct order. A sweep across 14 model families and 6 feature sets confirmed that the feature family, not the model, governs the outcome. Shapley attribution reduces exactly to two coefficients on named physiological quantities, so the pipeline from sensor to decision is auditable end to end.

The limits are bounded by measurement and by sample size rather than by the physiology: eleven participants support a resolvability claim rather than a deployment claim; movement and misregistration cannot be excluded by the current instrumentation and account for roughly half of the point estimate on a within-data motion surrogate; and the temporal statistic cannot be assigned to a named vasomotor band because the sampling interval sits below the Nyquist limit for every such band. The next study is a

better instrumented one, not a larger one, and the Discussion sets out the instrumentation and protocol changes that would tighten the uncertainty of the recovered signal and permit a formal Type A uncertainty budget in the sense of the Guide to the Expression of Uncertainty in Measurement.

Methods

Design, participants and protocol

A within-subjects repeated-measures design was used, so that every comparison reported below is paired and each participant serves as their own control; this removes between-person variation in resting vasomotor tone, which is the largest single source of variance in facial thermal data, and it is the reason a sample of eleven can support exact inference on the within-person contrast. Eleven healthy adults took part (7 men and 4 women; mean age 26.9 years, standard deviation 1.9, range 23 to 29). All gave written informed consent; none reported a cardiovascular, dermatological or neurological condition, and none was taking medication known to affect vascular tone. Participants were asked to avoid caffeine and vigorous exercise beforehand. The study was approved by the Institutional Ethics Committee of the Indian Institute of Technology Roorkee (approval EC/NEW/INST/2025/2056), was conducted in accordance with the Declaration of Helsinki, and all experiments were performed in accordance with the ethical standards of that committee and national research guidelines.

Sessions were run in a climate-controlled chamber maintained at 24°C plus or minus 1 degree, illuminated by approximately 150 lux of artificial light and 300 lux of daylight. Participants were seated at a fixed workstation with the thermal camera mounted frontally at 0.7 metres from the face at seated eye height. Room temperature and relative humidity were held constant across conditions and across participants, following the recommendations for human infrared thermography of Fernandez-Cuevas et al. ⁵ and the reporting guidelines subsequently derived from a systematic review of 315 facial thermography studies by Stanic and Gersak ⁶.

Each participant attended a single session, shown in Fig. 1a. A 15-minute briefing and acclimatisation period preceded any measurement, followed by an initial affect rating and a 10-minute silent reading task that served as the resting thermal baseline; acclimatisation and baseline together occupy the first 25 minutes of the visit, within the 15 to 20 minutes generally required for facial skin temperature to stabilise in a controlled environment. Two task blocks followed in randomised order (five participants received the stress task first, six the boredom task first). Each block lasted 15 minutes and had the same three-part structure, shown in Fig. 1b: two minutes of on-screen training on the task, eight minutes of the task itself, and five minutes of post-task questionnaires. The stress condition used the Montreal Imaging Stress Task, in which mental arithmetic is presented under a manipulated time limit alongside a display comparing the participant's performance against a purported average ²⁹; the task belongs to the same family of standardised psychosocial stressors as the Trier Social Stress Test ³⁰. The boredom condition used a visual memory game selected to be repetitive and undemanding. Both were administered in Inquisit 7. Thermal recording ran continuously through each block.

Subjective and thermal measures

Subjective assessments were administered at the end of the reading baseline and immediately after each task, and comprised four instruments. The Circumplex Model of Affect ³¹ captured valence and arousal on a two-dimensional bipolar grid from minus 2 to plus 2 on each axis. Direct task-state items rated the intended manipulation on a five-point Likert scale, covering perceived stress and time pressure after the stress task and monotony and boredom after the boredom task. An abbreviated NASA Task Load Index ³² assessed mental demand, time pressure, perceived success, effort and frustration on a bipolar verbal scale from minus 3 to plus 3. General-state items rated current stress, alertness and comfort with the room lighting on five-point scales, the lighting item serving as a control for environmental confounds.

Facial thermal data were acquired with a thermal infrared camera operating between 7.5 and 13 micrometres with a thermal sensitivity of plus or minus 0.05°C. Full-face images were captured once every 30 seconds, an effective sampling rate of approximately 0.033 Hz. Facial landmark detection on co-registered visible-spectrum images defined 24 anatomical regions, shown in Fig. 2, following a canonical mean-shape model of the kind used in modular thermal face-tracking pipelines ⁸. Six regions lie on the midline (chin, forehead, glabella, lips, moustache area, nose); the remaining eighteen form nine bilateral pairs (upper and lower cheeks, eye hoods, eyebrows, temples, irises, and inner, outer and inner-corner eye contours). The set covers both the external carotid territory, which includes the nose, cheeks, lips and temples, and the internal carotid and ophthalmic territory, which includes the forehead, glabella and periorbital regions ^{3,4}.

Feature extraction

For each region in each frame, six pixel statistics were recorded: mean, median, minimum, maximum, standard deviation and temperature range. This yields 144 instantaneous features per frame, dividing by definition, and in advance of any analysis, into two families: 96 describe absolute temperature (mean, median, minimum, maximum) and 48 describe spatial heterogeneity within a region (standard deviation, temperature range). This partition is the operational form of the physical prediction set out in the Introduction.

Absolute facial temperature varies substantially between individuals and between regions within an individual, so raw values are not comparable across participants. Each value was converted to a standard score relative to the same participant's own reading baseline for the same region and pixel statistic, using all available baseline frames (5 to 21 per participant, median 13). This procedure removes any static offset specific to a participant or region, including any fixed optical asymmetry from directional daylight in the chamber, and is the standard remedy for the between-participant variability documented by Ioannou et al. ²⁰.

Temporal features describe how each instantaneous series evolves rather than where it sits. Sixteen temporal summary statistics were computed for every series, following the general framework for scalable time-series feature extraction developed by Christ et al.³³: mean, standard deviation, minimum, maximum, range, median, skewness, kurtosis, linear slope, rate of change, lag-one autocorrelation, early-to-late difference, peak change, time to peak, approximate entropy and permutation entropy. The primary temporal feature is the rate of change, defined as the mean absolute difference between consecutive frames of a series and averaged across all 144 series to give one global measure of thermal volatility per session. It is unchanged by the addition of any constant, so it is insensitive to slow drift and to any fixed offset, and it responds to fluctuation of any sign, so opposing local changes reinforce rather than cancel, which is the behaviour expected under asynchronous anastomosis closure. Approximate entropy and permutation entropy^{34,35} were computed but excluded from all reported analyses, because at 16 samples both degenerate.

Analysis used frames 5 to 20 of each block, which is 16 frames spanning minutes 2 to 10; the window is fixed on protocol grounds rather than selected from the data. The first two minutes are the on-screen training period, identical in structure across conditions; the eight minutes that follow are the induction period proper. The window was available in all 22 sessions. One consequence of fixing the frame count rather than the elapsed duration requires disclosure: the 16 analysed frames span 7.58 minutes on average during stress and 8.56 minutes during the boredom task, because the mean interval between frames was 30.3 seconds and 34.3 seconds respectively; the Results report the primary result under four treatments of this asymmetry, including two in which both conditions are resampled onto a single common time grid.

Statistical analysis and classification

All primary tests are paired and within-participant. Rather than assume a distributional form at a sample size where that assumption cannot be checked, the null distribution for each test was constructed by exact enumeration: for eleven participants there are 2 to the power 11, that is 2,048, possible assignments of sign to the eleven participant-level differences, and every one was evaluated. The resulting P value is exact and distribution-free. One property governs how every P value in this paper should be read: the smallest attainable two-sided P value is 2 divided by 2,048, which is 0.00098, so a result reported at $P = 0.001$ is as significant as a design of this size can express and is not evidence of an unusually large effect. Effect sizes are reported as d_z , the mean paired difference divided by the standard deviation of that difference³⁶.

Analyses are separated into a small primary set, comprising the global omnibus test on the rate of change, the directional test on the rate of change, the phase-specificity contrast and the classification test, corrected by the Holm procedure across those four, and a larger exploratory screen of all 144 instantaneous features, corrected by the Benjamini-Hochberg procedure³⁷. Two disclosures belong here rather than in the limitations. The primary set was designated after the exploratory screen had been examined, and the rate-of-change measure was selected from fourteen candidate temporal summaries

computed on these same data; the primary set is therefore offered as a specification for replication rather than as an out-of-sample confirmation, and the Discussion prices the selection explicitly. Second, the limits of the exploratory screen can be stated in advance and arithmetically: at eleven participants, Holm correction across 144 tests requires a standardised effect above 1.60, and 144 multiplied by the P value floor of 0.00098 is 0.141, so no single feature in that screen can reach a corrected 0.05 by any route. Any feature-wise result from the screen is consequently descriptive and reported as such. To test whether a feature family carries any condition information without incurring a multiplicity penalty, a single omnibus statistic was computed for each family, the sum of squared t statistics across all 144 features, referred to the same exact sign-flip null.

Classification used leave-one-subject-out validation; on each of eleven folds both sessions belonging to one participant were held out and the model was fitted on the remaining ten, with feature scaling, feature selection and hyperparameter choice performed inside the training folds only through an inner leave-one-subject-out loop. No information from a held-out participant entered the model that predicted it. Subject-level validation of this kind is necessary to avoid the optimistic bias that frame-level or record-level splitting introduces in small samples²¹. Two models are reported. The first is a fixed two-feature logistic regression using the global rate of change and the within-region standard deviation of the lower left cheek, with the regularisation constant held at 1; this pair was identified from the full-data ranking and then held fixed across folds, and it is described as fixed rather than pre-specified for that reason, with the cost of the choice quantified in the Discussion. The second is a fully automatic pipeline in which the inner loop selects both the number of features and the regularisation constant, and which consequently uses no knowledge derived from the complete dataset. Absolute accuracy over 22 sessions is given with a Wilson 95 per cent confidence interval³⁸; forced-choice accuracy over 11 participants asks whether the stress session of a given participant received a higher predicted probability than that same participant's boredom session, and is the more appropriate summary because the 22 sessions form eleven clusters of two. Two control analyses accompany the main results: a permutation test that re-runs the entire fitting procedure under all 2,048 within-participant label flips, and a jackknife that drops each participant in turn and repeats the whole pipeline. A sweep of 14 model families crossed with 6 feature sets was also run under the same nested procedure and is reported in Results as a selection-cost calculation rather than as a result; the families were logistic regression under three penalties, ridge, linear discriminant analysis, naive Bayes, k -nearest neighbours, support vector machines with linear and radial kernels⁵¹, random forests, extremely randomised trees, gradient boosting and a multi-layer perceptron. For attribution, the Shapley additive explanation of a feature in a linear model reduces exactly to the product of its fitted coefficient and its centred value^{39,40}, so for the two-feature model the coefficients are reported directly. Because the applied question is which facial region an instrumented system should track, regions were additionally ranked on the rate of change and five region subsets defined by vascular anatomy rather than by that ranking were evaluated, so that no selection cost attaches to them.

Software and reproducibility

Two data corrections were applied before analysis: a twelve-hour clock wrap in one participant's boredom-condition timestamps, and one duplicated timestamp in another participant's stress-condition record. No participant and no session was excluded, and no baseline was repaired or substituted.

Declarations

Code availability

Analyses were implemented in Python 3 using NumPy, SciPy, pandas and scikit-learn. The full analysis code, including feature extraction, exact sign-flip inference and leave-one-subject-out classification, is available from the corresponding author on request and will be deposited in a public repository with a persistent DOI on acceptance.

Data availability

The facial thermal recordings and derived feature matrices that support the findings of this study are available from the corresponding author, but restrictions apply to the availability of these data, which were collected under ethical approval that does not permit public release of identifiable biometric data, and so are not publicly available. The data are, however, available upon reasonable request and with the permission of the Institutional Ethics Committee of the Indian Institute of Technology Roorkee. The analysis code is available from the corresponding author on request and will be deposited in a public repository with a persistent DOI upon acceptance.

Author contributions

S.R. conceived the study and designed the experimental protocol under the supervision of P.S.C. and R.E. S.R. collected the data, developed the analysis pipeline, performed the statistical analysis and classification, and prepared the figures. S.R. drafted the manuscript. P.S.C. and R.E. supervised the study, contributed to the interpretation of results, and reviewed and edited the manuscript. All authors approved the final version.

Funding

This work was supported by the Prime Minister's Research Fellowship, Ministry of Education, Government of India. The funder had no role in study design, data collection, analysis, interpretation, or the decision to submit for publication.

Competing interests

The authors declare no competing interests.

Declaration of generative AI use

During the preparation of this work, the author(s) used Claude (Anthropic) for grammar checking and for general consistency checking of the manuscript text. The author(s) reviewed and edited the output as

needed and take full responsibility for the content of the published article. No AI tool was used to generate data, to perform or design any analysis, or to draw any scientific conclusion.

References

1. Calderon-Uribe, S., Morales-Hernandez, L. A., Guzman-Sandoval, V. M. & Dominguez-Trejo, B. Cruz-Albarran, I. A. Emotion detection based on infrared thermography: A review of machine learning and deep learning algorithms. *Infrared Phys. Technol.* **145**, 105669 (2025).
2. Perpetuini, D., Filippini, C., Cardone, D. & Merla, A. An overview of thermal infrared imaging-based screenings during pandemic emergencies. *Int. J. Environ. Res. Public Health.* **18** (6), 3286 (2021).
3. Charkoudian, N. Skin blood flow in adult human thermoregulation: How it works, when it does not, and why. *Mayo Clinic Proceedings*, 78(5), 603–612 (2003).
4. Kuraoka, K. & Nakamura, K. Facial temperature and pupil size as indicators of internal state in primates. *Neurosci. Res.* **175**, 25–37 (2022).
5. Fernandez-Cuevas, I. et al. Classification of factors influencing the use of infrared thermography in humans: A review. *Infrared Phys. Technol.* **71**, 28–55 (2015).
6. Stanic, V. & Gersak, G. Facial thermal imaging: A systematic review with guidelines and measurement uncertainty estimation. *Measurement* **242**, 115879 (2025).
7. Tian, X., Fang, L. & Liu, W. The influencing factors and an error correction method of the use of infrared thermography in human facial skin temperature. *Build. Environ.* **244**, 110805 (2023).
8. Kopaczka, M., Breuer, L., Schock, J. & Merhof, D. A modular system for detection, tracking and analysis of human faces in thermal infrared recordings. *Sensors* **19** (19), 4135 (2019).
9. Sonkusare, S. et al. Detecting changes in facial temperature induced by a sudden auditory stimulus based on deep learning-assisted face tracking. *Sci. Rep.* **9**, 4729 (2019).
10. Charkoudian, N. Mechanisms and modifiers of reflex induced cutaneous vasodilation and vasoconstriction in humans. *J. Appl. Physiol.* **109** (4), 1221–1228 (2010).
11. Ulrich-Lai, Y. M. & Herman, J. P. Neural regulation of endocrine and autonomic stress responses. *Nat. Rev. Neurosci.* **10** (6), 397–409 (2009).
12. Kataoka, N., Shima, Y., Nakajima, K. & Nakamura, K. A central master driver of psychosocial stress responses in the rat. *Science* **367** (6482), 1105–1112 (2020).
13. Walløe, L. Arterio-venous anastomoses in the human skin and their role in temperature control. *Temperature* **3** (1), 92–103 (2016).
14. Nakayama, K., Goto, S., Kuraoka, K. & Nakamura, K. Decrease in nasal temperature of rhesus monkeys (*Macaca mulatta*) in negative emotional state. *Physiol. Behav.* **84** (5), 783–790 (2005).
15. Kuraoka, K. & Nakamura, K. The use of nasal skin temperature measurements in studying emotion in macaque monkeys. *Physiol. Behav.* **102** (3–4), 347–355 (2011).
16. Engert, V. et al. Exploring the use of thermal infrared imaging in human stress research. *PLOS ONE.* **9** (3), e90782 (2014).

17. Bini, G., Hagbarth, K. E., Hynninen, P. & Wallin, B. G. Thermoregulatory and rhythm-generating mechanisms governing the sudomotor and vasoconstrictor outflow in human cutaneous nerves. *J. Physiol.* **306**, 537–552 (1980).
18. Nayak, S., Nagesh, B., Routray, A. & Sarma, M. A human-computer interaction framework for emotion recognition through time-series thermal video sequences. *Comput. Electr. Eng.* **93**, 107280 (2021).
19. Filippini, C. et al. Automated affective computing based on bio-signals analysis and deep learning approach. *Sensors* **22** (5), 1789 (2022).
20. Ioannou, S., Gallese, V. & Merla, A. Thermal infrared imaging in psychophysiology: Potentialities and limits. *Psychophysiology* **51** (10), 951–963 (2014).
21. Vabalas, A., Gowen, E., Poliakoff, E. & Casson, A. J. Machine learning algorithm validation with a limited sample size. *PLOS ONE*. **14** (11), e0224365 (2019).
22. Merrifield, C. & Danckert, J. Characterizing the psychophysiological signature of boredom. *Exp. Brain Res.* **232** (2), 481–491 (2014).
23. Danckert, J. & Merrifield, C. Boredom, sustained attention and the default mode network. *Exp. Brain Res.* **236** (9), 2507–2518 (2018).
24. Van Tilburg, W. A. P. & Igou, E. R. On boredom: Lack of challenge and meaning as distinct boredom experiences. *Motivation Emot.* **36** (2), 181–194 (2012).
25. Pattyn, N., Neyt, X., Henderickx, D. & Soetens, E. Psychophysiological investigation of vigilance decrement. *Biol. Psychol.* **78** (3), 256–267 (2008).
26. Ioannou, S. et al. The autonomic signature of guilt in children: a thermal infrared imaging study. *PLoS ONE*. **8** (11), e79440 (2013).
27. Veltman, J. A. & Vos, W. K. Facial temperature as a measure of mental workload. In Proceedings of the 13th International Symposium on Aviation Psychology. Wright State University (2005).
28. Stemberger, J., Allison, R. S. & Schnell, T. Thermal imaging as a way to classify cognitive workload. In Proceedings of the 2010 Canadian Conference on Computer and Robot Vision (pp. 231–238). IEEE (2010).
29. Dedovic, K. et al. The Montreal Imaging Stress Task: Using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain. *J. Psychiatry Neurosci.* **30** (5), 319–325 (2005).
30. Kirschbaum, C., Pirke, K. M. & Hellhammer, D. H. The Trier Social Stress Test: A tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology* **28** (1–2), 76–81 (1993).
31. Russell, J. A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **39** (6), 1161–1178 (1980).
32. Hart, S. G. & Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (Eds.), *Human Mental Workload* (pp. 139–183). North-Holland (1988).

33. Christ, M., Braun, N., Neuffer, J. & Kempa-Liehr, A. W. Time series feature extraction on basis of scalable hypothesis tests (tsfresh - a Python package). *Neurocomputing* **307**, 72–77 (2018).
34. Richman, J. S. & Moorman, J. R. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology: Heart and Circulatory Physiology*, 278(6), H2039 to H2049 (2000).
35. Bandt, C. & Pompe, B. Permutation entropy: A natural complexity measure for time series. *Phys. Rev. Lett.* **88** (17), 174102 (2002).
36. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* 2nd edn (Lawrence Erlbaum Associates, 1988).
37. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Stat. Soc. Ser. B.* **57** (1), 289–300 (1995).
38. Wilson, E. B. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.* **22** (158), 209–212 (1927).
39. Shapley, L. S. A value for n-person games. In (eds Kuhn, H. W. & Tucker, A. W.) *Contributions to the Theory of Games II* (307–317). Princeton University Press (1953).
40. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. *Adv. Neural. Inf. Process. Syst.* **30**, 4765–4774 (2017).
41. Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* **180(Part A)**, 68–77 (2018).
42. Prasad, B. R. & Chandana, B. S. Human face emotions recognition from thermal images using DenseNet. *Int. J. Electr. Comput. Eng. Syst.* **14** (2), 155–167 (2023).
43. Assiri, B. & Hossain, M. A. Face emotion recognition based on infrared thermal imagery by applying machine learning and parallelism. *Math. Biosci. Eng.* **20** (1), 913–929 (2023).
44. Cardone, D., Pinti, P. & Merla, A. Thermal infrared imaging-based computational psychophysiology for psychometrics. *Computational and Mathematical Methods in Medicine*, 984353 (2015). (2015).
45. Pavlidis, I. et al. Fast by nature: How stress patterns define human experience and performance in dexterous tasks. *Sci. Rep.* **2**, 305 (2012).
46. Shastri, D., Papadakis, M., Tsiamyrtzis, P., Bass, B. & Pavlidis, I. Perinasal imaging of physiological stress and its affective potential. *IEEE Trans. Affect. Comput.* **3** (3), 366–378 (2012).
47. Kosonogov, V. et al. Facial thermal variations: A new marker of emotional arousal. *PLOS ONE*. **12** (9), e0183592 (2017).
48. Stefanovska, A., Bracic, M. & Kvernmo, H. D. Wavelet analysis of oscillations in the peripheral blood circulation measured by laser Doppler technique. *IEEE Trans. Biomed. Eng.* **46** (10), 1230–1239 (1999).
49. Julien, C. The enigma of Mayer waves: Facts and models. *Cardiovascular. Res.* **70** (1), 12–21 (2006).
50. Dickerson, S. S. & Kemeny, M. E. Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research. *Psychol. Bull.* **130** (3), 355–391 (2004).

51. Vapnik, V. *The Nature of Statistical Learning Theory* (Springer, 1995).

52. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14** (5), 365–376 (2013).

Figures

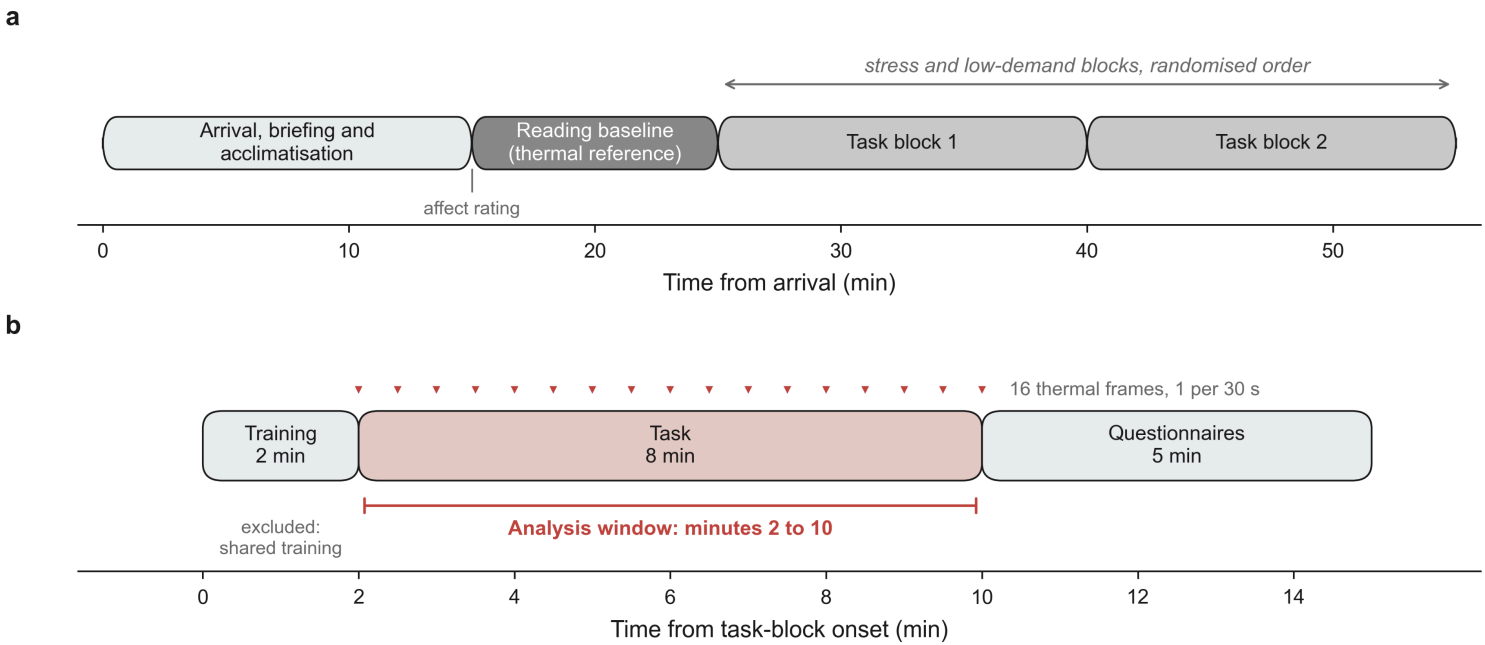


Figure 1

Experimental protocol. (a) Session timeline: acclimatisation, reading baseline, then two task blocks in randomised order. (b) One task block: 2 min training, 8 min task and 5 min questionnaires; the 16 analysed frames span minutes 2 to 10.

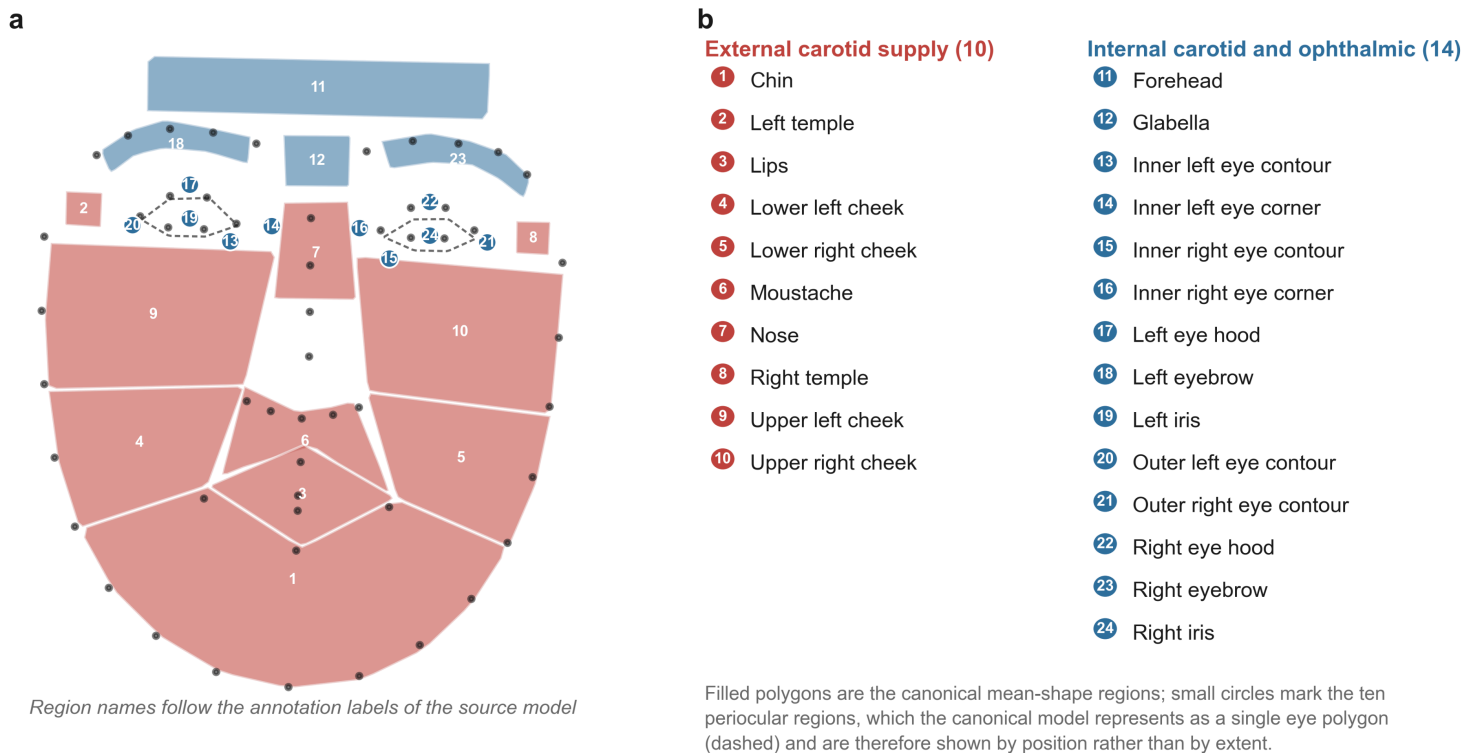


Figure 2

The 24 facial regions of interest. (a) Region geometry on the canonical mean-shape model. (b) Key, grouped by arterial supply. External carotid regions are subject to sympathetic vasoconstriction; the ophthalmic territory is relatively protected.

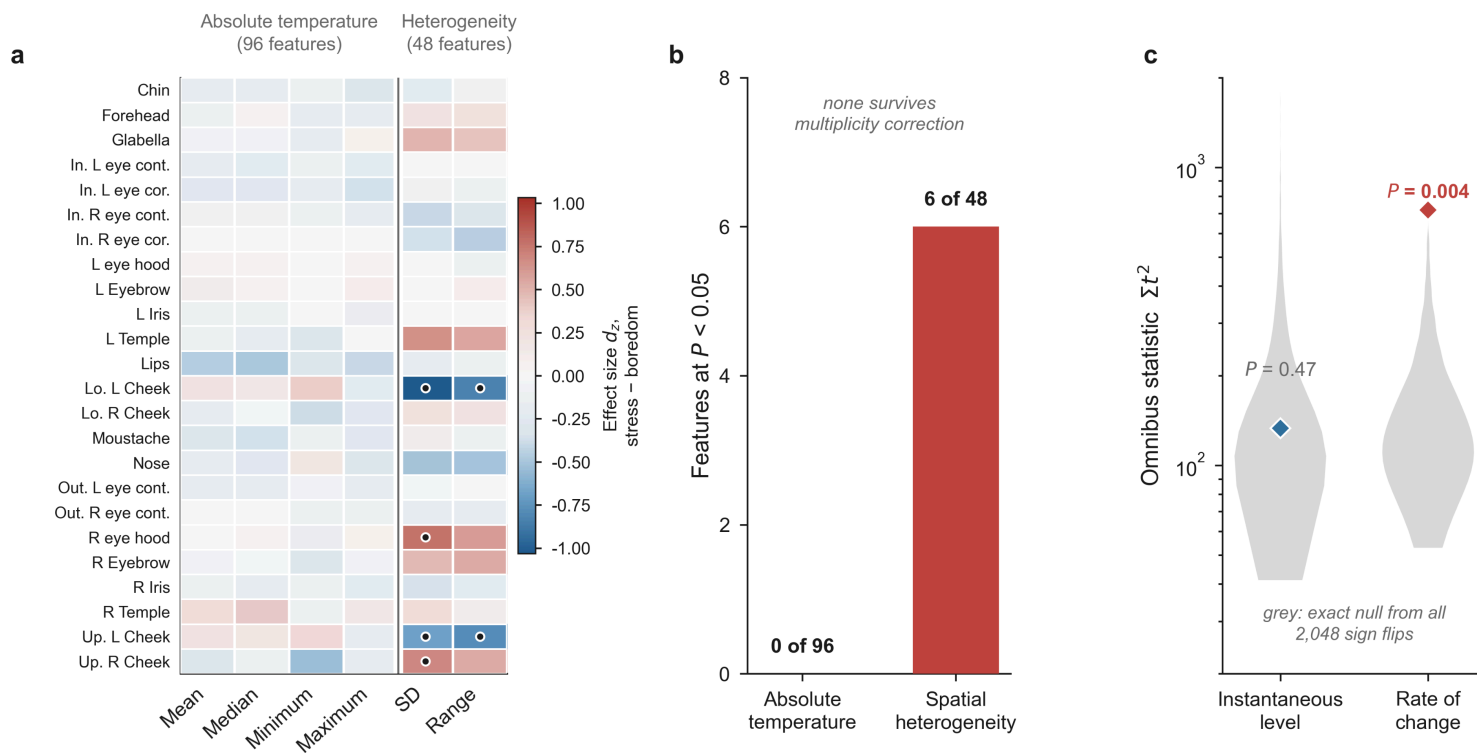


Figure 3

Instantaneous features carry no level effect. (a) Effect size for each of the 144 region-by-statistic contrasts; dots mark exact $P < 0.05$. (b) Counts of significant features by family. (c) Omnibus statistic for each family against its exact sign-flip null.

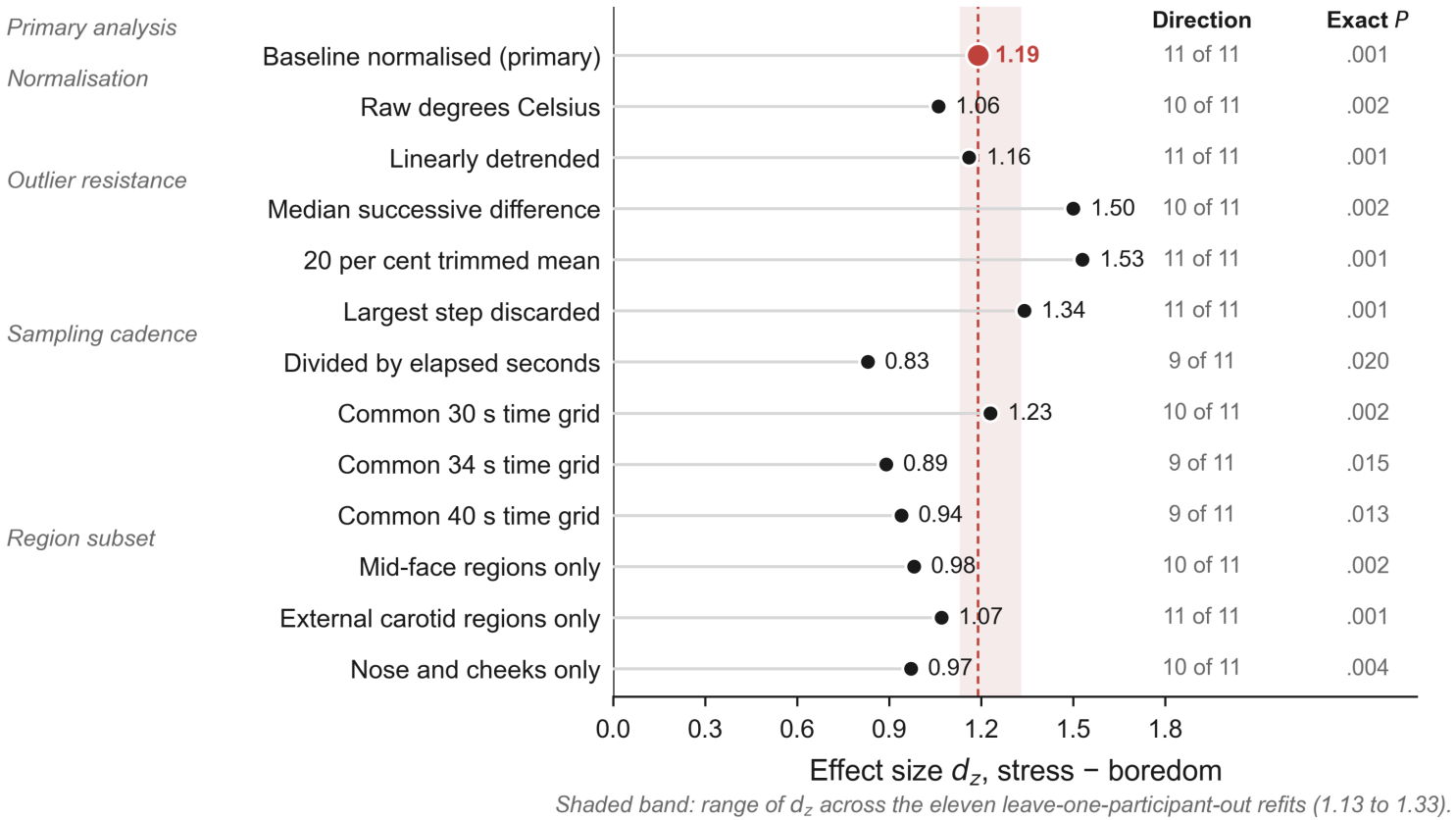


Figure 4

The rate-of-change effect under every treatment applied to it. Points are grouped by the challenge each treatment represents. The primary analysis is highlighted; the shaded band gives the range across eleven leave-one-participant-out refits.

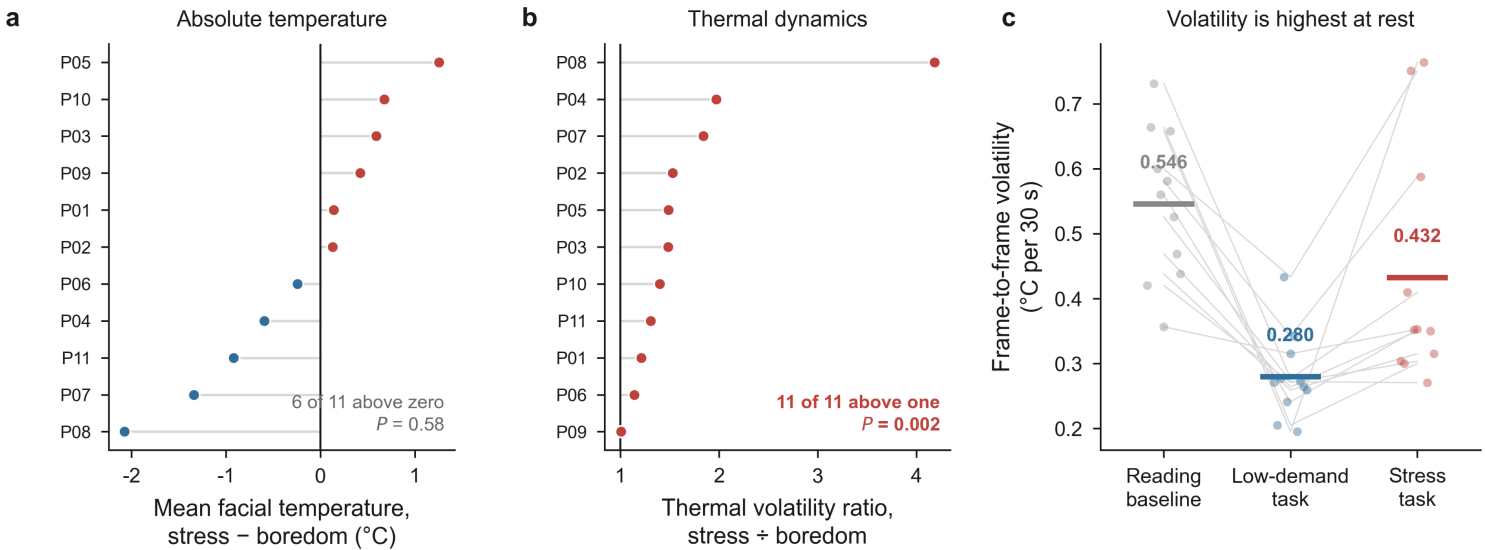


Figure 5

Dynamics and level compared on scale-free footings. (a) Per-participant temperature change. (b) Per-participant volatility ratio. (c) Absolute volatility in the three states, showing that volatility during both tasks is lower than during the reading baseline.

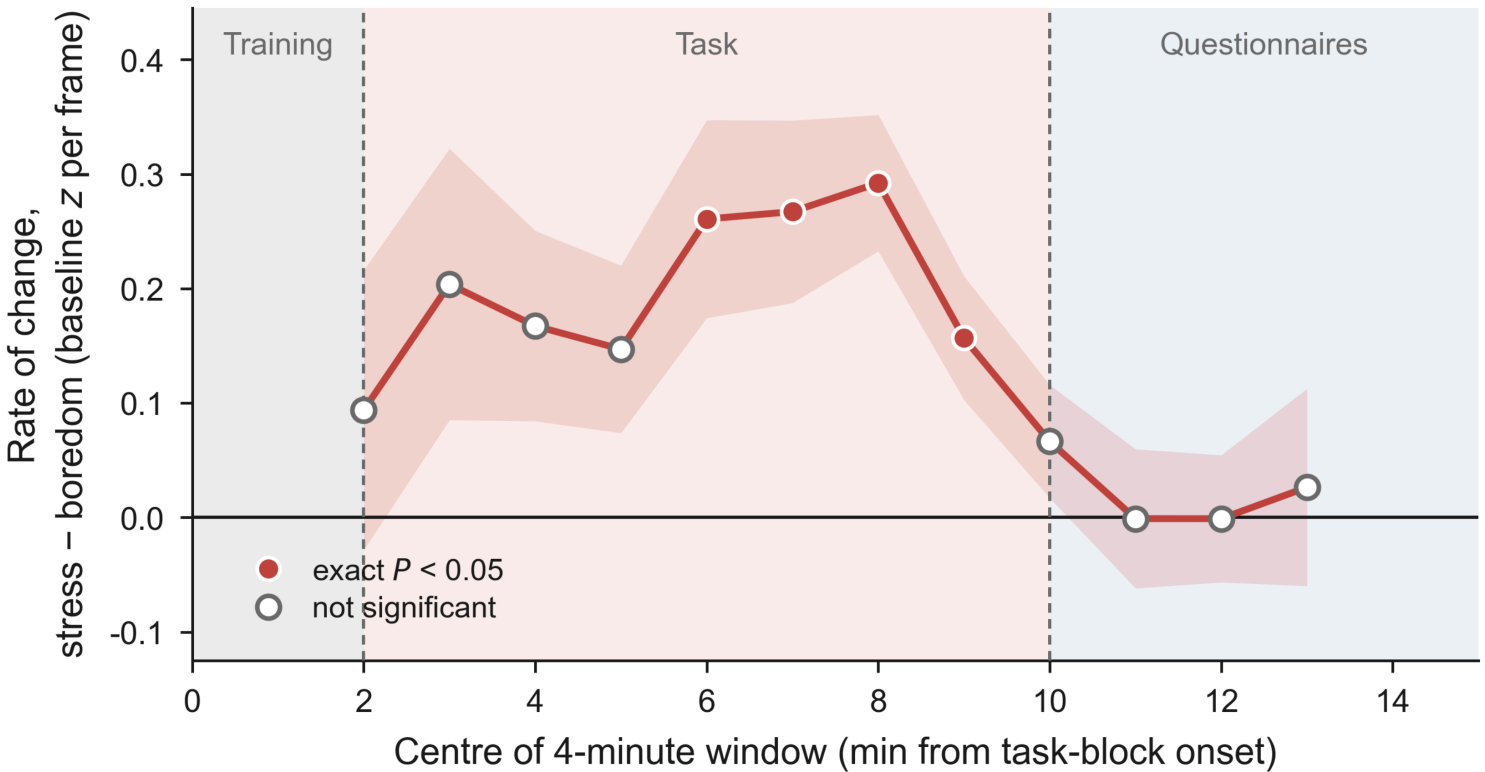


Figure 6

The difference is confined to the task period. Rate-of-change difference in sliding four-minute windows across the 15-minute block. Filled markers indicate exact $P < 0.05$; the band is one standard error across participants.

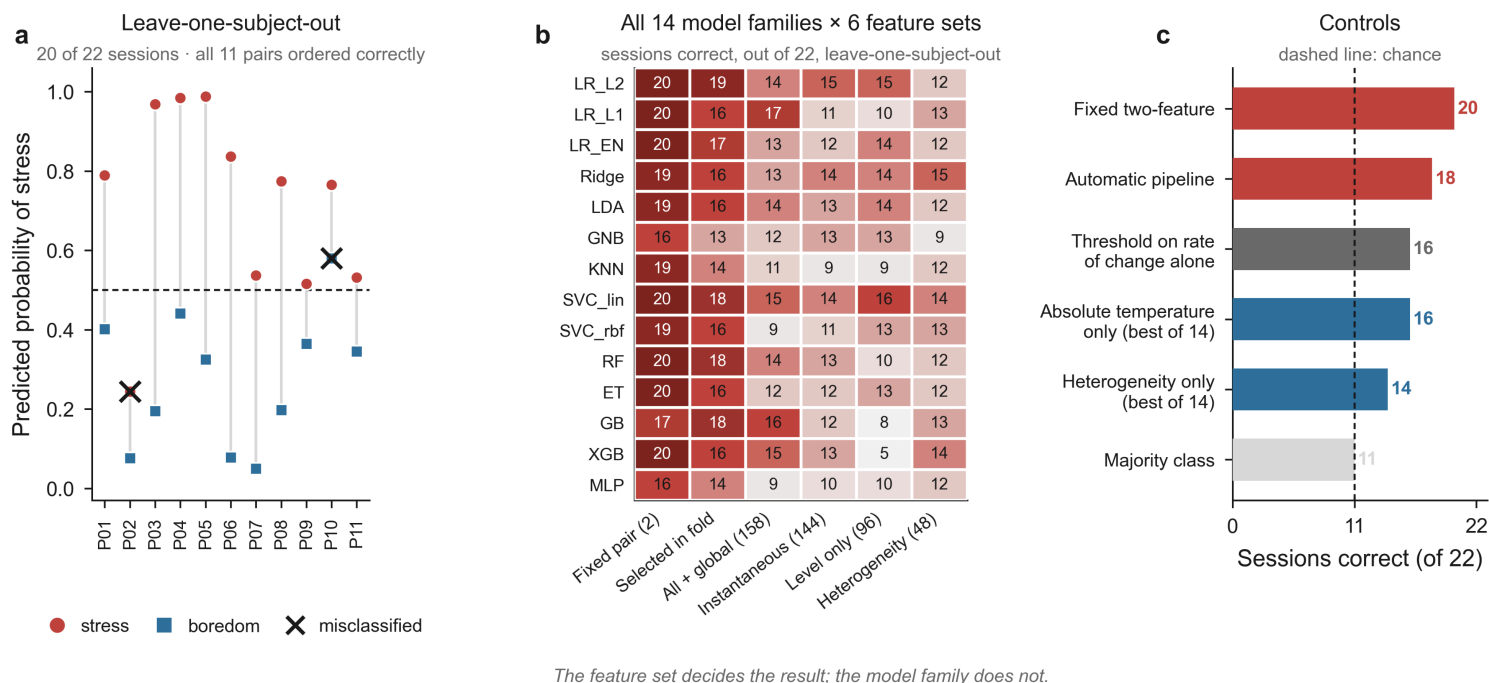


Figure 7

Discriminating stress from boredom under leave-one-subject-out validation. (a) Predicted probability of stress for each session. (b) Accuracy for every combination of 14 model families and 6 feature sets. (c) Control analyses, with the majority-class floor marked.

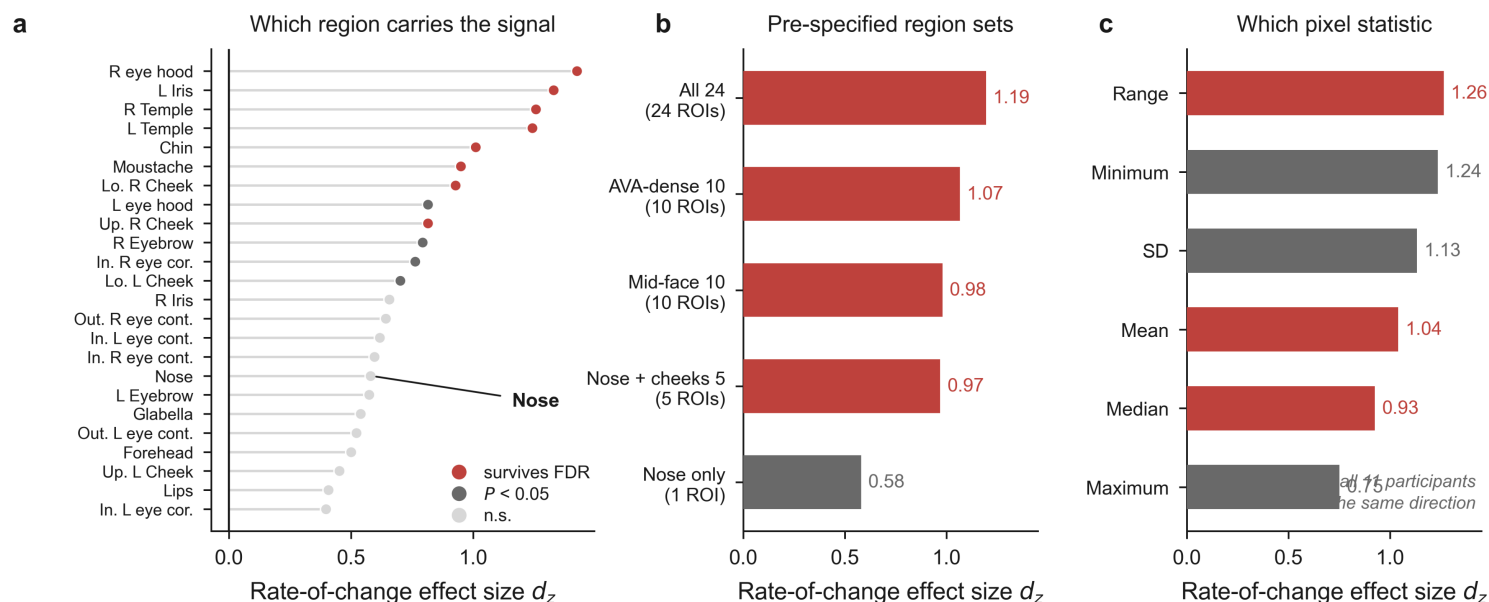


Figure 8

What an instrumented system would need to record. (a) All 24 regions ranked by rate-of-change effect size, with the nose marked. (b) Region subsets defined by vascular anatomy. (c) Effect size computed on each of the six pixel statistics separately.