Original Research Article

# Detecting Cognitive Stress in Knowledge Work Using Keyboard and Mouse Interaction Behavior

Venkat Nithin Porala

*Archbishop Mitty High School, 5000 Mitty Way, San Jose, CA 95129, United States*

## ABSTRACT

Cognitive stress is a common obstacle in knowledge-requiring workspaces and is recognized to influence both performance and motor control. The most established stress-detection systems require complex hardware including physiological and wearable sensors that detect the following factors: heart rate, electrodermal activity, respiratory activity, skin temperature, and pupil/eye measures. Systems have rarely been created with the purpose of stress detection through behavioral interaction signals alone. This study explores this lacuna by investigating whether cognitive stress can be detected from keyboard and mouse behavior alone. Keyboard and mouse interaction offers a minimal-sensing alternative. Using the publicly available SWELL-KW dataset, the study extracts nine interaction features (five keyboard and four mouse features) over one-minute intervals. Logistic regression and random forest classifiers are trained then evaluate the nine features under a leave-one-participant-out cross-validation against a majority-class baseline. The accuracy under simple behaviors is below that of complex multi-modal systems, which is the cost of looking at the feasibility of the model. The behavioral signal is genuine but weak: pooled ROC-AUC reaches 0.58-0.61 (above the 0.50 chance level) and PR-AUC 0.66-0.70 (above the 0.615 prevalence baseline), and macro-F1 0.54-0.58. Both modalities performed comparably, with total mouse-movement distance as the strongest single predictor. Combining the modalities produces a statistical gain for both models. Aggregating predictions to the task-block level improved discrimination (ROC-AUC up to 0.72). Behavior-only stress detection is feasible and non-intrusive, but currently provides weak, highly person-dependent accuracy.

**Keywords:** cognitive stress; stress detection; keyboard and mouse dynamics; behavioral sensing; human-computer interaction; minimal-sensor monitoring; machine learning; SWELL-KW

## INTRODUCTION

Cognitive stress is the mental strain that builds when cognitive demands, such as intense thinking, attention, or decision-making, exceeds an individual's capacity (1). This mental overload can affect many noticeable aspects of a person, some of which include the following: well-being, productivity, and motor control (2). In a workplace, cognitive stress is built from information overload, time pressure, unclear explanations, etc. The buildup of cognitive stress leads to the deviation of certain measurable metrics. Specifically, this paper investigates how cognitive stress can be measured through the evaluation of motor control. Motor control is the ability of the nervous system to plan and coordinate movements to perform everyday tasks (2). Due to the fact that most modern knowledge work requires continuous keyboard

**Corresponding author:** Venkat Nithin Porala, E-mail: nithinporala@gmail.com.

**Accepted** July 2, 2026
https://doi.org/10.70251/HYJR2348.44106117

and mouse interaction, behaviors that depend on precise motor control, changes in coordination from stress may give surface in how people type and move a cursor. Prior research has affirmed that increased cognitive stress affects neural functioning in ways that impair attention and coordination (1). Detecting these behavioral fluctuations early is valuable for sustaining efficient performance and well-being in work environments (1).

Current approaches to effectively detecting cognitive stress rely on heart rate, electrocardiography, electrodermal activity, and respiration patterns; these signals are proven to have a reliable correlation between cognitive stress and the body's responses (3, 4). In particular, heart-rate variability in particular has a well-documented meta-analytic association with stress (5). Though these signals are accurate in controlled environments, they are challenging to employ in professional workspaces; they require complex and costly hardware, require per-device calibration, reduce user compliance, raise privacy concerns, and are difficult to sustain over long periods of time(4). Similarly, multimodal approaches are effective, but their dependence on sensors and machinery shifts the focus toward behavioral signals that naturally emerge during everyday work.

Keyboard and mouse interaction have been used in prior stress and workload research, but primarily as supporting inputs alongside physiological sensors rather than as signals evaluated in isolation, leaving their independent predictive power under-examined (6, 7). Furthermore, existing studies tend to emphasize classification accuracy rather than examining the tradeoff between sensing complexity and real-world feasibility (6). For monitoring systems to be mass deployable, feasibility is a primary design constraint. As a result, there exists a gap in this field where systems that experimentally perform well in controlled environments cannot realistically be deployed in professional workspaces.

From this gap, evaluating parameters through the lens of feasibility is important to minimizing the complexity while maintaining moderate accuracy in the process. Building off of previous studies (6), this research investigates whether behavioral signals, keyboard and mouse features, which can be collected from computers without any complex sensors, alone are sufficient to identify whether an individual is stressed or non-stressed. This study explores the extent to which data collected from keyboards and mice can be used to conclude the cognitive state of a working person. This paper contributes to stress-detection in multiple ways: evaluating the effectiveness of lightweight machine learning models(logistic regression and random forest), low-complexity classifiers that can be trained and run with modest computational resources such as a low memory CPU with no specialized hardware, on behavioral motor-control data, examining the tradeoff between sensor complexity and real-world detection feasibility, and demonstrating that practical, non-intrusive sensing approaches can provide a modest but meaningful stress signal in everyday knowledge-work environments.

## LITERATURE REIVEW

### Cognitive Stress and Its Effect

Cognitive Stress, as mentioned in the introduction, refers to the mental strain that results from the exceeded demand for an individual's cognitive resources (2). This mental state commonly activates in situations that require an elongated time period of concentration, time pressure, multitasking, etc (2). Existing research in cognitive psychology shows that attentional control and task performance can be heavily degraded by increased mental effort (2). According to the limited-capacity view of attention, mental resources are finite, and when the limit is exceeded, performance and quality of motor execution decline (2). In addition, these cognitive stress levels also correlate with motor behavior by slowing reaction time and reducing the coordination of the motor functions of an individual. (1) proved that an extended period of this stressed state can ultimately negatively impact productivity, learning, and overall well-being of the individual under stress, making it a pressing issue in modern work environments that require long periods of focus (1). However, the early detection of stress can greatly impact these long-term effects, as individuals can improve their cognitive state by taking a break and relaxing (1).

### Physiological Approaches to Stress Detection

In the past, many studies have used physiological signals to detect stress as a result of their strong association with the human body's response to stress. The signals that were used were many: heart rate variability, electrocardiography, electrodermal activity, and respiration. Studies by (3) and (4) demonstrated that these signals have a strong correlation to the stress level of an individual and are highly reliable, particularly in controlled lab environments. For example, a meta-analysis of heart-rate variability studies found a consistent association between reduced HRV and elevated stress (5). These successful connections explain why physiological signals are the primary parameters in stress detection research. However, these signals rely on complex equipment that collects data through wearable sensors, which can limit their practicality for everyday workspace detection. Furthermore, these monitoring systems often involve calibration procedures for each machine, limiting the feasibility for many people. These limitations spark the exploration for alternative signals to predict the cognitive state of a person with an acceptable or nearly the level of success that sensing allows.

**Behavioral Stress Indicators in Human-Computer Interaction**

Human behavior as they perform tasks that require motor function reflects underlying cognitive and emotional states, which include stress (2). Changes often occur in behavior as a person is under stress, as tasks that require focus demand cognitive attention, which naturally decreases over time. Prior research in human-computer interaction has shown that keyboard dynamics can vary when an individual is under stress compared to the same person under a non-stress state (6, 7). Similarly, mouse interaction patterns are also motor functions that require focus, evidently making them susceptible to change when under stress; specifically, the motor coordination and decision-making change under stress (2). The main property that keyboard and mouse activity provides is that it is continuous and passive during computer use, meaning it can be captured without additional hardware or interruption between the user. As a result, computer interaction represents a promising source of data to evaluate if a person is stressed or not.

**Multimodal Stress Detection Systems**

Multimodal stress detection systems help improve the performance of detection by combining multiple sources of information that capture complementary aspects of the stress response (4). In professional environments and office-related studies, these systems integrate physiological signals with behavioral interaction data, facial expression, posture information, etc (6). (4) demonstrated that combining multiple modalities can help improve the accuracy of detection systems; this would be through the capture of both internal physiological responses and external behaviors. Due to this, these approaches are used frequently in experiments that use stress detection systems. However, multimodal systems also introduce more complexity in data analysis, synchronizing multiple data sources. The level of computational power needed for these approaches can exceed what is feasible for casual workspace use. They also introduce the need for multiple sensors, which increases the intrusiveness and has largely limited their use to controlled settings. While multimodal methods provide performance, these constraints limit their suitability in everyday knowledge-work settings.

**Stress Datasets for Knowledge Work: SWELL-KW**

One widely used benchmark for studying stress and workload is the SWELL-DW(SWELL - Knowledge Work) dataset that presents data in office-based knowledge work environments (6). The dataset comprises recordings from 25 participants, each contributing roughly three hours of data while performing realistic knowledge-work tasks, like writing reports, preparing presentations, reading email, and searching for information (6). The data was collected from participants who were performing computer-based tasks under different conditions: both neutral(no-stress) and two stress-inducing conditions(email interruptions and time pressure) (6). In the interruption condition, participants received multiple emails, some of which were relevant and required action and others irrelevant; in the time-pressure condition, the time to complete comparable tasks was reduced to two-thirds of the neutral duration (6). This dataset is designed for a multimodal analysis of stress, pairing computer-interaction logs with physiological signals(heart rate, heart-rate variability, and skin conductance), facial-expression video, and body-posture data (6). Ground-truth labels were established from a validated self-report questionnaire on task load, mental effort, emotion, and perceived stress after each condition (6). This dataset has been heavily used in stress-detection research; for example, studies that use a multimodal approach to maximize the accuracy of the detection (6). In these works, keyboard and mouse interaction data are typically used as supporting features rather than as the primary source of information. As a result, although SWELL-KW contains rich behavioral interaction data, the independent contribution of keyboard and mouse signals to stress detection remains insufficiently explored. The conditions in the research reflect experimentally induced workload rather than self-identified stress, a distinction that depends on how the results should be interpreted (6). This makes the dataset well-suited for investigating minimal-sensor approaches that focus on behavior-only stress detection in realistic work settings.

## METHODS AND MATERIALS

### Data Description

This study employs the SWELL-KW (6) dataset, a benchmark designed to study stress and workload in office-based knowledge work environments. The publicly available dataset was obtained through an approved SWELL-KW collection distributed through the DANS repository (6). No new human data was collected for this study, an individual institutional review-board approval was therefore not applicable. The data was collected from 25 participants performing realistic computer-based tasks under controlled experimental conditions (6). Each participant completed three tasks in four condition types: neutral, interruption, time-pressure, and relaxation. Block 1 was the neutral condition, while both Blocks 2 and 3 contained the two stress conditions, which were in a counterbalanced order across all 25 participants. The SWELL-KW dataset includes multiple sensing modalities, such as keyboard interaction logs, mouse movement data, physiological signals, posture measurements, and facial expression data. In this study, only keyboard and mouse interaction data was used. All physiological, postural, and facial modalities were intentionally excluded to evaluate behavior-only stress detection under minimal sensing constraints, reflecting a deployment scenario where additional hardware is impractical. To be specific, the analysis in the study drew on two files: the raw uLog events (75 XML files, one per participant per block) for feature engineering and the official per-minute computer-interaction feature file as an authoritative source of condition labels.

### Behavioral Data Preprocessing

Raw keyboard and mouse interaction logs were extracted from the SWELL-KW dataset and processed before feature extraction. Each uLog file records a list of timestamped events. Every event carries an event type(keyboard, mouse, or other), an action, and a sub-millisecond UTC timestamp (6). The logs record a single event per keystroke and mouse coordinates, rather than continuously, only when an interaction(click, drag, wheel) is made (6). Due to the fact that the features are derived from UTC-stamped raw logs while the condition labels were taken from a feature file recorded in Netherlands local time, the two were aligned at the minute level using a daylight-saving-aware conversion to Amsterdam time (6). Raw events were combined in fixed, non-overlapping one-minute windows aligned to clock minutes; a window was retained only if it contained at least one logged event, so inactive minutes(relaxed) produce no row. Data from each participant were kept separate to prevent cross-participant contamination during training and evaluation. Feature standardization was fit only on each training fold (Section F), and missing values were imputed using training-fold statistics only, so no information leaked from test to train.

### Feature Extraction

Behavioral features were extracted from the processed interaction data to capture changes associated with cognitive stress, guided by prior human-computer-interaction work and a preference for interpretability. Nine features were computed per one-minute window, five keyboard features and four mouse features, including: typing speed, mean inter-key delay, inter-key delay (SD), backspace/error rate, typing pause frequency, mean cursor speed, total movement distance, click rate, and mouse idle proportion.

Table 1 includes each feature's precise operational definition. Several thresholds were fixed in advance: a keyboard "pause" is an inter-key gap exceeding 2.0 seconds; "error" keys are Backspace and Delete; typing speed is measured in character keystrokes per minute; and idle proportion is the fraction of a window's sixty one-second bins containing no logged event.

Edge cases were handled explicitly; when a window contained fewer than two keyboard or mouse events, the affected timing features were recorded as missing rather than imputed with zero. Two features used in some prior analysis were deliberately excluded because the logs cannot support them: key-press duration, which requires key-release events the logs do not contain, and cursor-movement smoothness, which requires a continuous cursor log.

### Stress Labeling Strategy

Stress detection was framed as a binary classification problem in which the interruption (I) and time-pressure (T) conditions were labeled stress (positive class) and the neutral (N) condition was labeled non-stress (negative class). Condition labels were taken from the official per-minute feature file rather than inferred from block number, which would be incorrect given the counterbalancing of the stress conditions. These labels reflect experimentally induced, task-related stress rather than clinically assessed stress, so the models distinguished stress-related task conditions instead of medical stress states. After dropping the relaxation logs, the analysis dataset comprised 2,653 labeled minutes

***Table 1.*** *Operational definitions of the five keyboard and four mouse features extracted per one-minute window from the SWELL-KW dataset. "Error" keys are Backspace and Delete. A typing "pause" is an inter-key gap that exceeds 2.0 seconds. Idle proportion is the fraction of a window's sixty one-second bins without logged events. These nine features are the model inputs throughout.*

| Feature | Modality | Operational definition |
|---|---|---|
| **Typing speed** | Keyboard | Count of character key events per 1-min window (chars/min) |
| **Inter-key delay (mean)** | Keyboard | Mean gap (s) between consecutive keyboard events in the window |
| **Inter-key delay (SD)** | Keyboard | SD of gaps (s) between consecutive keyboard events |
| **Backspace/error rate** | Keyboard | Fraction of keyboard events that are Backspace/Delete |
| **Pause frequency** | Keyboard | Number of inter-key gaps exceeding 2.0 s |
| **Mean cursor speed** | Mouse | Sum of move distance / sum of inter-event time over mouse events (px/s) |
| **Total movement distance** | Mouse | Sum of per-event mouse relative distance (px) |
| **Click rate** | Mouse | Count of mouse click events per window |
| **Idle proportion** | Mouse | Fraction of the 60 one-second bins containing no logged event |

from 25 participants: 1,022 non-stress and 1,631 stress, with a minute-level stress prevalence of 0.6148 (Table 2). Every participant contributed both classes, and the modest sample size is acknowledged as a limitation in the Discussion section.

### Machine Learning Models

A total of three models were evaluated under identical cross-validation splits. First, a trivial baseline (constant prediction of the most frequent training class) provides a no-information reference. Next, logistic regression and random forest were chosen for interpretability, low computational cost, and suitability for standard hardware. No deep-learning or sequence models were used in this study.

All the hyperparameters were fixed beforehand and never tuned on held-out folds (Table 3). The logistic regression model used the lbfgs solver with C = 1.0, max_iter = 1000, and balanced class weights (Table 3). The random forest model used 300 trees, unrestricted depth, square-root feature sampling, and balanced class weights (Table 3). Each model was wrapped in a pipeline

***Table 2.*** *Composition of the analysis dataset after preprocessing (25 participants). Stress combines the interruption(I) and time-pressure(T) conditions. Non-stress is the neutral(N) condition. Relaxed minutes were excluded. Counts are given at the one-minute level (2,653 labeled minutes with a minute-level prevalence of 0.615) and the task-block level (75 participant sessions with a block-level prevalence of 0.667).*

| Quantity | Value |
|---|---|
| **Participants** | 25 |
| **Labeled minutes (samples)** | 2653 |
| **non-stress (N) minutes** | 1022 |
| **stress (I+T) minutes** | 1631 |
| **Minute-level stress prevalence** | 0.6148 |
| **Per-participant stress minutes (min/median/max)** | 46 / 68 / 70 |
| **Per-participant non-stress minutes (min/median/max)** | 32 / 42 / 44 |
| **Block structure** | B1=N (non-stress); B2,B3=I,T stress (counterbalanced) |
| **Block-level units (PP x block)** | 75 (25 N, 50 I/T) |
| **Block-level stress prevalence** | 0.667 |

***Table 3.*** *Full specifications of the three models and their preprocessing in scikit-learn 1.8.0 (numpy 2.4.4, pandas 3.0.2, and scipy 1.17.1, under Python 3.12). The trivial baseline predicts the most frequent training class (no-skill reference). All hyperparameters were fixed beforehand. To prevent leakage, imputation and standardization were fit on each fold's training partition.*

| Component | Specification |
|---|---|
| **Trivial baseline** | Most-frequent-class predictor (constant); used as no-skill reference |
| **Logistic regression** | solver = lbfgs; C = 1.0; max_iter = 1000; class_weight = balanced; random_state = 42 |
| **Logistic regression preprocessing** | Per-fold median imputation -> standardization (zero mean, unit variance), fit on training partition only |
| **Random forest** | n_estimators = 300; max_depth = None; max_features = sqrt; class_weight = balanced; random_state = 42; n_jobs = -1 |
| **Random forest preprocessing** | Per-fold median imputation, fit on training partition only |
| **Global seed** | 42 (all stochastic components) |

performing per-fold median imputations fit on each training partition. The implementation used scikit-learn 1.8.0, with numpy 2.4.4, pandas 3.0.2, and scipy 1.17.1, under Python 3.12 and a single global random seed of 42 fixed for all stochastic components (Table 3). Balanced classes were applied during training due to the fact that the classes are imbalanced; furthermore, by a sensitivity check, it was confirmed that removing class weighting changed aggregate discrimination by less than 0.01.

### Evaluation Protocol

Leave-one-participant-out cross-validation was used to assess generalization to unseen individuals; this was implemented as a leave-one-group-out scheme grouped by participant. Of the 25 folds, all minutes from 24 participants formed the training set and all minutes from the single held-out participant formed the test set, so each participant served as the test set exactly once. Because grouping is by participant, no minute from a test participant ever appears in training, preventing the model from exploiting participant-specific idiosyncrasies and precluding adjacent-minute leakage across the train and test boundary. Threshold-independent discrimination metrics were seen as primary because the class imbalance made threshold-based metrics misleading, which was the area under the ROC curve (ROC-AUC) and the area under the precision-recall curve (PR-AUC). Macro-averaged F1 and accuracy were reported as secondary metrics. AUCs were summarized both per participant (mean, SD, min, median, max across the 25 held-out participants) and as a single pooled out-of-fold value over all minutes for each model. No-skill references were reported alongside: ROC-AUC = 0.50 and PR-AUC = 0.615 (the positive-class prevalence). The parser was validated against the official feature file before analysis: per-minute character and keystroke counts correlated strongly with the official columns ($r = 0.89$ and $r = 0.95$), confirming correct minute-level alignment.

### Supplementary Analyses and Design Constraints

This study was intentionally designed under constrained sensing conditions: only keyboard and mouse data were used, and the models were chosen to be lightweight rather than optimized for maximal accuracy. Three supplementary analyses were conducted on the same splits, seeds, and preprocessing.

First, every model was evaluated under keyboard-only, mouse-only, and combined configurations to isolate each feature's contribution. Second, out-of-fold per-minute predictions were aggregated to the task-block level (one participant-by-condition session; 75 blocks, prevalence 0.667), with block discrimination reported as the headline and label-based metrics reported under an explicit caveat that the fixed 0.5 threshold is mismatched to the 0.667 block mix. Third, permutation importance (20 repeats per held-out fold, scored by ROC-AUC and averaged across folds) was computed for both models, cross-checked against logistic-regression coefficients, to identify the features driving cross-participant discrimination. Pairwise model and modality comparisons were tested across the 25 folds with the Wilcoxon signed-rank test (complemented by a paired t-test) and interpreted as evidence about direction and plausibility rather than as definitive significance claims, given the 25 paired observations and modest effect sizes.

## RESULTS

After preprocessing, the complete SWELL-KW interaction logs yielded 2,653 labeled one-minute windows from all 25 participants in the dataset, 1,022 non-stress (N) and 1,631 stress (T and I) minutes, which came to a minute-level prevalence of 0.615. The data comprises 75 participant sessions, 25 non-stress and 50 stress(a block prevalence of 0.667). All three models were evaluated with leave-one-participant-out cross-validation. Threshold-independent discrimination metrics, ROC-AUC and PR-AUC, are reported as primary, with macro-averaged F1 and accuracy as secondary; no-skill references are ROC-AUC = 0.50 and PR-AUC = 0.615. Any differences are interpreted against the expected variability between participants.

### Cross-Participant Detection Performance

Table 4 records the minute level performance on all nine features from both the keyboard and mouse data. The trivial baseline stays parallel with the no-skill references of ROC-AUC 0.500 and PR-AUC 0.615; this confirms the metric implementation. Next, both random forest and logistic regression classifiers exceeded no-skill, but only with a marginal improvement. Logistic regression reached a pooled out-of-fold ROC-AUC of 0.614 and PR-AUC of 0.697. Random forest reached a pooled out-of-fold ROC-AUC of 0.588 and PR-AUC of 0.663. For logistic regression, the per-participant mean ROC-AUC was 0.636 and 0.588 for random forest. Accuracy stayed around the 0.615 majority baseline for both classifiers (logistic regression 0.586, random forest 0.602), and macro-F1 (logistic regression 0.579, random forest 0.538). This showed a marginal to moderate separability. Logistic regression attained a higher AUC than random forest. The pooled ROC and precision-recall curves, depicted in Figures 1 and 2, show that both models tracked above the chance diagonal and the prevalence line across operating points.

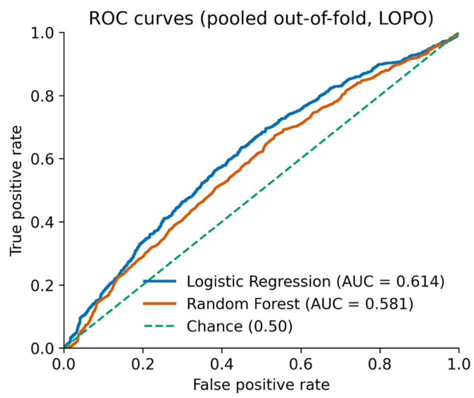


***Figure 1.*** *Pooled out-of-fold ROC curves under LOPO for logistic regression and random forest on all the nine features with the chance diagonal (ROC-AUC = 0.500) for reference.*

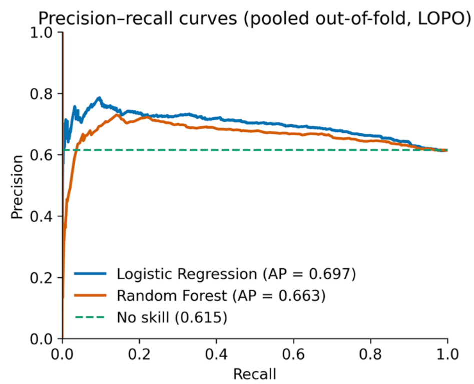


***Figure 2.*** *Pooled out-of-fold precision-recall curves under LOPO for logistic regression and random forest, with the prevalence baseline (PR-AUC = 0.615) for reference.*

***Table 4.*** *Minute-level performance of the three models on all the nine features under leave-one-participant-out (LOPO) cross-validation. ROC-AUC (the area under the receiver-operating characteristic curve) and PR-AUC (the area under the precision-recall curve) are primary and macro-F1 and accuracy are secondary. AUCs are reported per participant ±SD across the 25 held-out participants and as a pooled out-of-fold value. ROC-AUC = 0.500, PR-AUC = 0.615 (prevalence).*

| Model | macro-F1 | Accuracy | ROC-AUC (per-pp) | ROC-AUC pooled | PR-AUC (per-pp) | PR-AUC pooled |
|---|---|---|---|---|---|---|
| **Trivial** | 0.381 | 0.615 | 0.500 ± 0.000 [0.500/0.500/0.500] | 0.500 | 0.613 ± 0.031 [0.523/0.622/0.683] | 0.615 |
| **LogReg** | 0.579 | 0.586 | 0.636 ± 0.142 [0.292/0.604/0.967] | 0.614 | 0.737 ± 0.096 [0.534/0.741/0.981] | 0.697 |
| **RandForest** | 0.538 | 0.602 | 0.588 ± 0.086 [0.437/0.607/0.748] | 0.581 | 0.691 ± 0.064 [0.581/0.691/0.796] | 0.663 |

### Contribution of Keyboard and Mouse Modalities

To isolate each modality, every model was retrained under 3 conditions: keyboard-only, mouse-only, and combined features (Table 5). Mouse-only features performed slightly higher, but not statistically, than the keyboard-only features. Combining both modalities produced the best results for each classifier; the gain for random forest and logistic regression over keyboard-only features was significant (random forest: +0.053 F1 and $p = 0.011$; logistic regression: +0.070 F1 and $p = 0.011$). When comparing the two models, it is crucial to recognize that model rank is metric-dependent; notably, logistic regression resulted in higher ROC-AUC and macro-F1, while random forest achieved higher positive-class (stress) F1. Mouse behavior carries stress-related information comparable to keyboard behavior, and the two modalities are complementary rather than redundant.

### Feature Importance

Feature importances were derived by analysis on the participant held out of each LOPO fold. Then they were scored by ROC-AUC and averaged over folds, with logistic regression permutation importance and standardized coefficients as cross-checks (Table 6, Figure 3). All three measures, total mouse movement distance and typing speed (which respectively ranked first and second), and error rate, pause frequency, and click rate (all three of which were near zero), agreed. This is consistent with the modality analysis. Fold-to-fold standard deviations equal or exceed the mean importances for several features, so although the direction is consistent across all three methods, the exact ordering below the top two features is not robust and is interpreted with corresponding caution.

***Table 5.*** *LOPO performance of logistic regression and random forest under keyboard-only, mouse-only, and combined sets. ROC-AUC and PR-AUC are primary and macro-F1 and accuracy are secondary.*

| Feature set | Model | ROC-AUC | PR-AUC | macro-F1 | Accuracy |
|---|---|---|---|---|---|
| **Keyboard** | Logistic Reg. | 0.557 | 0.663 | 0.538 | 0.539 |
| **Keyboard** | Random Forest | 0.531 | 0.642 | 0.508 | 0.555 |
| **Mouse** | Logistic Reg. | 0.605 | 0.701 | 0.576 | 0.584 |
| **Mouse** | Random Forest | 0.557 | 0.659 | 0.529 | 0.577 |
| **Combined** | Logistic Reg. | 0.614 | 0.697 | 0.579 | 0.586 |
| **Combined** | Random Forest | 0.581 | 0.663 | 0.538 | 0.602 |

***Table 6.*** *Feature importance, ranked by random forest permutation importance on the held-out participant of each LOPO fold, with logistic-regression permutation importance and standardized coefficients as cross-checks.*

| Feature | Modality | RF perm. imp. | LR perm. imp. | LR std. coef. |
|---|---|---|---|---|
| **Total movement distance** | Mouse | +0.0422 ± 0.0664 | +0.1039 ± 0.1301 | +0.4412 ± 0.0293 |
| **Typing speed** | Keyboard | +0.0227 ± 0.0453 | +0.0601 ± 0.0863 | +0.3243 ± 0.0402 |
| **Inter-key delay (mean)** | Keyboard | +0.0202 ± 0.0228 | +0.0155 ± 0.0278 | +0.1842 ± 0.0263 |
| **Inter-key delay (SD)** | Keyboard | +0.0108 ± 0.0288 | -0.0013 ± 0.0044 | +0.0299 ± 0.0172 |
| **Idle proportion** | Mouse | +0.0039 ± 0.0291 | +0.0048 ± 0.0283 | -0.1193 ± 0.0437 |
| **Backspace/error rate** | Keyboard | -0.0005 ± 0.0156 | -0.0022 ± 0.0045 | +0.0170 ± 0.0129 |
| **Click rate** | Mouse | -0.0011 ± 0.0216 | -0.0058 ± 0.0097 | -0.0389 ± 0.0435 |
| **Mean cursor speed** | Mouse | -0.0036 ± 0.0310 | +0.0002 ± 0.0011 | +0.0856 ± 0.0039 |
| **Pause frequency** | Keyboard | -0.0071 ± 0.0109 | +0.0005 ± 0.0164 | -0.0760 ± 0.0181 |

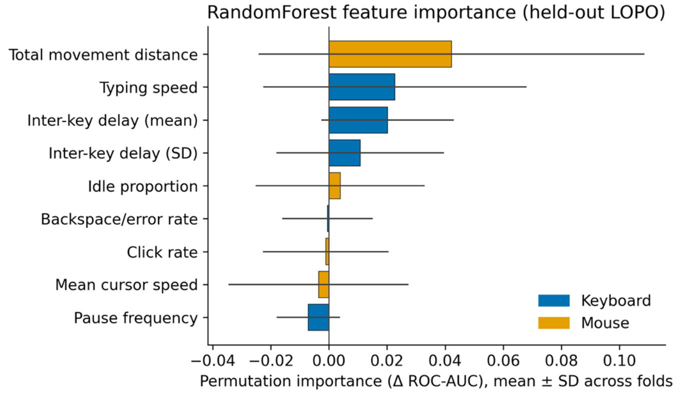


***Figure 3****. Random Forest permutation importance ranked for the nine features. Bars are colored by modality.*

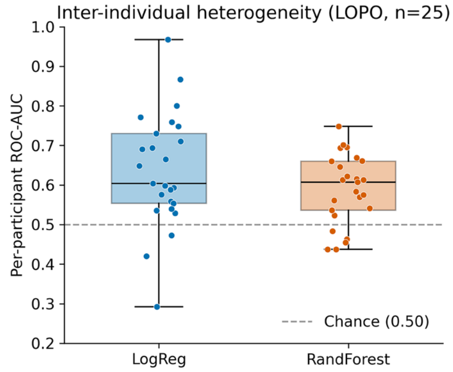


***Figure 4****. Per-participant ROC-AUC under LOPO for logistic regression and random forest, shown as a boxplot with individual points and a chance like at 0.50.*

**Inter-Individual Heterogeneity**

Performance between participants varied across individuals as seen in Table 7 and Figure 4. Per-participant ROC-AUC for logistic regression spanned 0.292 to 0.967 (mean 0.636, SD 0.142, IQR 0.176), and for random forest 0.437 to 0.748 (mean 0.588, SD 0.086, IQR 0.124). Most participants performed better than just chance (88% logistic regression, 80% random forest), and roughly half exceeded 0.60 (52% for both). This range shows that the indication of behavioral stress is based on the individual instead of being uniform. As each of the 25 participants performs for a heavily restricted period of time, these estimates are unstable and are reported descriptively rather than as precise individual performance.

**Block-Level (Per-Session) Discrimination**

The out-of-fold per-minute predictions were aggregated to the entire set because in a real-world knowledge workspace, a per-session decision is a better representative than a per-minute one. Logistic-regression ROC-AUC rose from 0.614 to 0.698 and PR-AUC from 0.697 to 0.812; random-forest ROC-AUC rose from 0.581 to 0.724 and PR-AUC from 0.663 to 0.844 (Table 8;

***Table 7.*** *Per-participant ROC-AUC distribution under LOPO (n=25) for logistic regression and random forest: mean, SD, min, median, max, interquartile range (IQR), and percentage of participants exceeding 0.50 and 0.60.*

| Model | Mean | SD | Min | Median | Max | IQR | % > 0.50 | % > 0.60 |
|---|---|---|---|---|---|---|---|---|
| **LogReg** | 0.636 | 0.142 | 0.292 | 0.604 | 0.967 | 0.176 | 88% | 52% |
| **RandForest** | 0.588 | 0.086 | 0.437 | 0.607 | 0.748 | 0.124 | 80% | 52% |

***Table 8.*** *Block Level v. minute-level discrimination under LOPO for both classifiers. Block-level aggregates each participant's per-minute out-of-fold scores within a condition session. Two labeling rules are shown. ROC-AUC and PR-AUC rise with the aggregation, while macro-F1 can fall.*

| Model | Level (rule) | ROC-AUC | PR-AUC | macro-F1 | Accuracy |
|---|---|---|---|---|---|
| **Logistic Reg.** | minute | 0.614 | 0.697 | 0.579 | 0.586 |
| **Logistic Reg.** | block (majority) | 0.698 | 0.812 | 0.639 | 0.653 |
| **Logistic Reg.** | block (mean-prob) | 0.698 | 0.812 | 0.655 | 0.667 |
| **Random Forest** | minute | 0.581 | 0.663 | 0.538 | 0.602 |
| **Random Forest** | block (majority) | 0.724 | 0.844 | 0.473 | 0.680 |
| **Random Forest** | block (mean-prob) | 0.724 | 0.844 | 0.395 | 0.653 |

Figures 5 and 6). The metrics based on labels, which were fixed at the 0.5, threshold did not improve uniformly; this is present in random-forest block macro-F1, which fell to 0.40-0.47 even as block ROC-AUC rose to 0.724. Figures 7 and 8 show the confusion matrices, which make the cause explicit. 23 of 25 non-stress blocks were labeled as stress because the fixed 0.5 threshold is mismatched to the 0.667 prevalence. This is a thresholding product and not a loss of discrimination. A system deployed on the block-level would require threshold recalibration. The block-level analysis in this study rests only on 75 units, so the corresponding AUCs are unstable and are also reported with that caveat.

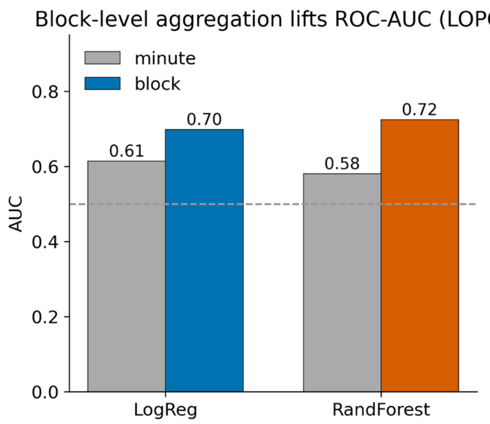


***Figure 5.*** *Block Level v. Minute Level ROC-AUC under LOPO for logistic regression and random forest, showing the gain from aggregating per-minute predictions to session level.*

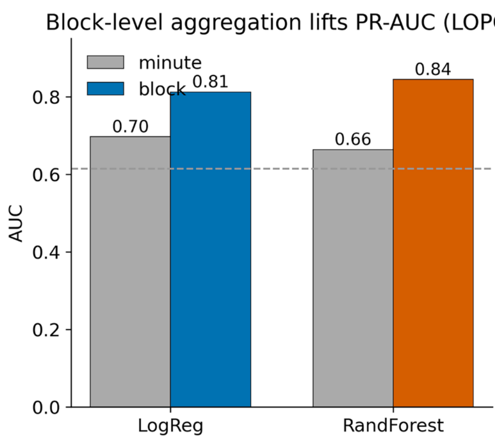


***Figure 6.*** *Block Level v. PR-AUC under LOPO for logistic regression and random forest, showing the gain from aggregating per-minute predictions to session level.*

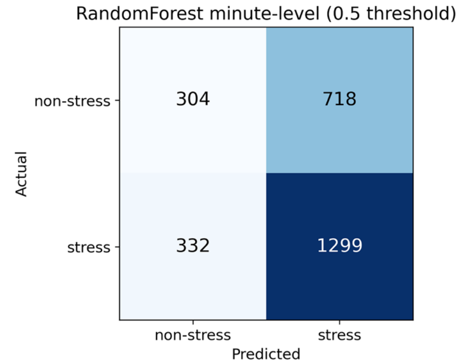


***Figure 7.*** *Minute-level confusion matrix for the random forest classifier under LOPO cross-validation. This uses the fixed 0.5 decision as a threshold. Rows are the actual class (non-stress and stress) and columns are the predicted class. Cells are minute counts (n = 2,653). In the 1,022 non-stress minutes, 718 were misclassified as stress. This shows the model's bias towards the stress class at this threshold.*

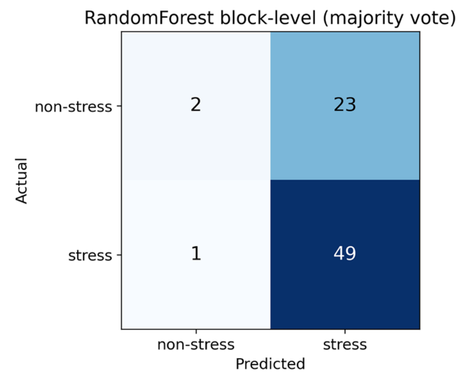


***Figure 8.*** *Block-level confusion matrix for the random forest classifier under LOPO cross-validation. This uses the majority-vote aggregation of predictions per minute within each participant by condition session (n = 75). Rows are the actual class (non-stress and stress) and columns are the predicted class. 23 out of 25 non-stress blocks were labeled stress.*

## DISCUSSION

To sum up, this paper experimented on whether only simple motor functions performed through keyboards and mice could distinguish the cognitive state of an individual. The finding is that a weak to moderate signal exists and that both modalities have the same level of contribution rather than dominating. In addition, discrimination improves when predictions are aggregated to the task level. Mouse-only models marginally performed better than keyboard-only models;, still, the difference between the performance was not significant. Combining modalities to train on nine features produced a small but significant gain from both classifiers (random forest +0.053 F1, $p = 0.011$; logistic regression +0.070 F1, $p = 0.011$). This shows that keyboard and mouse features help each other and add unique relevance.

Across all three measures, random forest permutation importance, logistic regression permutation importance, and standardized coefficients, total mouse movement distance was the strongest predictor, with typing speed as second. Error-related typing behavior and pause frequency ranked near zero, so error and hesitation behaviors are not clear markers of stress in this cross-participant setting. Fold-to-fold standard deviations equalled or exceeded the mean importances below the top two features, so while the direction is consistent across methods, the precise ranking of lower features is not robust.

Both models behaved, though differently, in a way that shows their structure rather than one being significantly better than the other. The models' ranks are metric-dependent, so neither model is outright superior. Logistic regression resulted in higher ROC-AUC and macro-F1, while random forest achieved higher positive-class (stress) F1. Both models use different strategies to give output: logistic regression assumes a linear relationship between features, whereas random forests can represent non-linear and threshold effects and higher-order interactions. As combining both modalities helped both models, this suggests that the complementary value of keyboard and mouse features is also reliably accessible to a linear model.

Another finding explains the level to which performance depends on the individual. Per-participant ROC-AUC ranged from 0.292 to 0.967 for logistic regression and 0.437 to 0.748 for random forest; this ranges from below chance (0.5) to almost perfect (1.0). This indicates that the behavioral depiction of stress depends from person-to-person, which means that a single population-level model will work better in some tests and worse in others. An opportunity for improvement from this finding is creating a model that is calibrated per user, as prediction performance depends on each individual's stress patterns.

Shifting predictions from the minute level to the task-block level significantly improved discrimination. Random forest ROC-AUC improved from 0.581 to 0.724, and PR-AUC went from 0.663 to 0.844. Similarly, logistic regression ROC-AUC grew from 0.614 to 0.698. This is consistent with the deployment reality that stress monitoring cares about a person's state over a work period rather than in any single minute, and that averaging over many minutes cancels much of the per-minute noise. This gain should be read as a promising direction rather than a precise estimate, however, as block-level metrics are computed over only 75 units, and the label-based metrics additionally require threshold recalibration to the block-level class prevalence.

As stated in the introduction, from a deployment perspective, behavior-only sensing has practical advantages as it requires no additional hardware, operates passively while also reducing the cost, intrusiveness, and privacy exposure that are commonly associated with physiological and wearable sensors. The LOPO protocol means the reported performance reflects generalization to genuinely unseen individuals rather than memorized per-person patterns. However, when taken together the minute-level performance shows that this approach is a better fit as a low-cost, scalable supporting signal, like as one input among others, rather than as a strongly reliable stress diagnostic.

Several limitations should be considered when interpreting these findings. First, the stress labels reflect experimentally induced workload rather than clinically assessed stress. This means that the conclusions are limited to stress aroused from tasks only. Furthermore, the controlled setting does not include the real unpredictability of modern knowledge workplaces. Second, the dataset contains a limited and relatively homogeneous group of 25 participants with similar occupational profiles, narrowing generalizability to only a few types of professions and interaction styles and reduces statistical power. The SWELL-KW dataset only includes 25 participants with 46 to 70 minutes of stress each, and per-participant estimates are unstable. A larger and more diverse participant group would more likely lead to more stable estimates and give a reliable performance. Third, the modeling relies on handcrafted features over fixed one-minute windows, which does not capture long-range temporal structure in behavior (how typing rhythm evolves over the course of a task). Finally, behavioral signals may be shaped by factors beyond task-induced stress, including fatigue, individual experience, and task familiarity, that this study does not separate from the stress conditions; because the neutral condition always occurred first, the condition is partially confounded with practice effects, despite counterbalancing.

From these limitations, there are many opportunities for future research that builds off of this study. One way a model might be able to perform more accurately is if it was personally calibrated towards each individual instead of depending on other people who show unique behavior. In a different path, better performance might be achieved through training by a wider and more diverse range of people who are studied in a less controlled setting, which simulates the real world. Finally, multimodal comparison studies that quantify exactly how much accuracy is traded for the gain in feasibility would help position behavior-only sensing precisely along the accuracy-feasibility spectrum.

## CONCLUSION

By limiting the analysis to only everyday computer interactions, this paper investigated whether stress during knowledge work could be predicted just through keyboard and mouse behavior. The study prioritized feasibility over reliability, and to do so, only lightweight machine-learning models were used. By using the leave-one-participant-out method, the results show that behavioral signals carry a genuine but weak signal to the cognitive state of an individual. This demonstrates that behavior-only stress detection is feasible, but it should be taken into consideration that it is not a reliable diagnostic.

Among the modalities, both keyboard and mouse features contributed about the same, with total mouse-movement arising as the strongest single predictor. Combining the two modalities provided noticeable improvement over just evaluating based on one type of feature. Also, performance varied across individuals, suggesting that, rather than minute-level, session-level decisions are more realistic in modern knowledge workspaces.

While behavior only models do not provide the accuracy of multimodal systems that are trained on a wider set of modalities, they offer substantial advantages in cost, privacy, ease of deployment, and may scale more readily. As a result, this approach is best seen as a nonintrusive, low-cost supporting signal for continuous stress monitoring rather than as a replacement for complex sensing. As potential solutions will likely depend on person-adapting models, validation on larger and more diverse datasets is imperative. Ultimately, this work contributes honest empirical evidence on what minimal-sensor, behavior-based stress detection can and cannot deliver, complementing existing research that emphasizes more complex sensing infrastructures.

## ACKNOWLEDGEMENTS

The author thanks Claudia D'Ettorre from University College London for the development of this research paper.

## CONFLICT OF INTEREST

The author declares no conflict of interest related to this work.

## REFERENCES

1. McEwen BS. Protective and damaging effects of stress mediators. *New England Journal of Medicine*. 1998; 338 (3): 171-179. https://doi.org/10.1056/NEJM199801153380307
2. Kahneman D. Attention and Effort. 1973. *Prentice-Hall*.
3. Healey JA & Picard RW. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*. 2005; 6 (2): 156-166. https://doi.org/10.1109/TITS.2005.848368
4. Alberdi A, Aztiria A & Basarab A. Towards an automatic early stress recognition system: A review. *IEEE Transactions on Affective Computing*. 2016; 7 (2): 99-115.
5. Kim HG, Cheon EJ, Bai DS, Lee YH & Koo BH. Stress and heart rate variability: A meta-analysis. *Psychiatry Investigation*. 2018; 15 (3): 235-245. https://doi.org/10.30773/pi.2017.08.17
6. Koldijk S, Neerincx MA, Kraaij W, et al. Detecting workload in office work environments using unobtrusive sensors. *Proceedings of the International Conference on Intelligent User Interfaces*. 2014.
7. Pepa L, Sabatelli A, Ciabattoni L, Monteriù A, Lamberti F & Morra L. Stress Detection in Computer Users From Keyboard and Mouse Dynamics. *IEEE Transactions on Consumer Electronics*. 2021; 67 (1): 12-19. doi:10.1109/TCE.2020.3045228