Understanding Behind the Smile of Emotion Workers: Detecting After-Call Stress in Call Agents

Duri Lee
School of Computing
Korea Advanced Institute of Science & Technology (KAIST)
Daejeon, Republic of Korea
durilee@kaist.ac.kr

Heejeong Lim
Graduate School of Data Science
Korea Advanced Institute of Science & Technology (KAIST)
Daejeon, Republic of Korea
hj.lim@kaist.ac.kr

Vedant Das Swain*
Tandon School of Engineering
New York University
New York City, New York, USA
v.das.swain@nyu.edu

Uichin Lee*
School of Computing
Korea Advanced Institute of Science & Technology (KAIST)›
Daejeon, Republic of Korea
uclee@kaist.ac.kr

Abstract

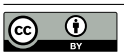
Call agents, a representative group of emotion workers, must manage emotions under constrained autonomy, yet workplace stress sensing has primarily centered on knowledge work. We ask how the task-aligned cycle of emotional labor, alternating customer interaction (CI) and non-customer interaction (nCI), shapes stress and how it manifests in data. We conducted a month-long in-the-wild formative mixed-methods study with professional call agents, collecting structured task logs, environmental and behavioral signals, and per-call stress self-reports, followed by semi-structured interviews. Task logs, used as a new sensor modality, were incorporated as primary sensing signals, and task-related features were extracted by respecting CI boundaries for modeling. Our results showed that a short 5-minute windowing approach was comparable to task-aligned windowing using multimodal sensors, with task-related features being considered the most important across all generalized models. Personalized models improved further and shifted importance toward diverse data sources, revealing individual differences in preparation patterns. Interviews support those findings, reveal key modelling challenges, and highlight potential benefits of semi-automated self-tracking. We discuss implications for timing interventions at breakpoints suited for work patterns, and ethically deploying stress support for emotion workers.

CCS Concepts

• **Human-centered computing** → *Empirical studies in ubiquitous and mobile computing*;

Keywords

emotion workers, stress sensing, multi-modality



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790803

ACM Reference Format:
Duri Lee, Heejeong Lim, Vedant Das Swain, and Uichin Lee. 2026. Understanding Behind the Smile of Emotion Workers: Detecting After-Call Stress in Call Agents. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3772318.3790803

1 Introduction

Emotion workers perform critical roles in the service industry (e.g., call centers, flight attendants, and sales persons) by managing emotions as part of their job functions [40]. They are required to express emotions as *display rule* needed for the organization while interacting with customers, a process referred to as emotional labor [32]. Call agents are prime examples of emotion workers, frequently performing emotional labor in their interactions with customers. As the initial point of contact in call centers, with a market size valued at USD 29.44 billion globally in 2024 [1], they play a pivotal role in facilitating customer interactions effectively. They are expected to resolve customer issues on time and ensure customer satisfaction [24], often under the constraints of limited autonomy and low agency to address issues independently [70]. These constraints, along with the emotional labor inherent in their roles, constitute a primary source of job stress.

Repeated exposure to emotionally demanding interactions with customers significantly contributes to the mental stress of call agents [99]. The sustained emotional labor gradually depletes a worker's mental resources, serving as a significant contributor to chronic stress. They endure dozens of emotionally taxing and mentally draining interactions daily, which exert detrimental effects on their physical and mental health over the long term [82]. A study conducted by Cornell University in 2017 reported that 87% of call agents experience high levels of stress [24], which leads to risks such as low job satisfaction and mental health issues (e.g., chronic stress and depression). Given the high level of stress and associated risks they face, effective strategies are necessary to protect them from excessive stress and enhance their well-being.

Data-driven stress monitoring systems have shown promise for supporting workplace well-being [11, 28, 78]. However, most studies have focused on knowledge workers and relied on digital traces (e.g., email logs [59], task logs [13], or generic wearable signals [42]) suited to unstructured, cognitively intensive tasks and time-fixed sampling, which fail to capture the episodic, interpersonal nature of call agents' work. The workflow of call agents fundamentally differs from that of knowledge workers. Their daily workflow alternates between non-customer interaction (nCI) and customer interaction (CI). nCI involves preparing for the next CI without display rules and CI requires adherence to explicit display rules, which means organizationally specified norms for emotional expression (e.g., maintain a warm and polite tone, express empathy for the caller's feelings, and avoid audible frustration). Therefore, employing static time windows, which is the standard approach in data-driven research for knowledge workers, risks misaligning these patterns and blurring interaction-specific stress signals. Moreover, these cycles generate structured task logs (e.g., call duration, inquiry records, and conversation topics) that differ from knowledge-work traces. These logs can provide valuable contextual information, such as when stress may arise, how long it lasts, and under what interactional conditions it occurs. Accordingly, we model after-call stress at the episode level, treating each call and its immediately adjacent nCI and CI as the unit of analysis, aligning features and labels to task boundaries, and elevating task logs to one of the sensing signals.

We focus on this task-based cycle because of its theoretical significance and its methodological advantages. Theoretically, this cycle is not merely a workflow but the fundamental structure of the agents' lived experience, reflecting the distinct 'front-stage' (CI) and 'back-stage' (nCI) roles of service occupations [31]. These transitions into different phases serve as powerful situational cues that trigger specific psychophysiological shifts [32], making the task cycle a meaningful unit of human experience, not just an operational one. Practically, these natural breakpoints offer practical methodological benefits. They provide ideal moments for precise, in-the-moment self-reports, thereby enhancing data validity; a meaningful unit for building analytical models that can capture context-specific patterns; and opportune moments for designing future, non-disruptive interventions [65]. Therefore, understanding the temporal and structural nature of call agents' stress by aligning the analysis with these task segments is an essential step toward designing effective strategies to protect them from excessive stress and enhance their well-being.

Building on the need to understand stress in this unique context, our study is guided by a central research question: *How does the task-aligned (CI and nCI) task cycle of emotional labor shape call agents' stress, and how are these dynamics reflected in call-center task logs, multimodal data, and call-aligned self-reports?* To answer this question, we conducted a month-long, in-the-wild field study that sequentially integrated quantitative and qualitative methods. First, we collected multimodal data (i.e., task logs, environmental signals, behavioral and physiological streams, daily baseline self-reports, and call-aligned self-reports of stress) and analyzed them with a unified modeling pipeline. Here, *task logs* denote operational metadata (e.g., call start/end timestamps, durations); we did not analyze call audio or transcripts. This pipeline systematically compared task-aligned vs. fixed windowing, data-source composition (source-wise ablations; with/without self-reports), and personalization schemes, using modeling as a lens to surface patterns. Second, we conducted post-study interviews and performed a thematic analysis to gain deeper, contextual insights into call agents' experiences. As a result, this study makes the following contributions.

- We use task logs as primary sensing signals and demonstrate that task-aligned (CI and nCI) features improve after-call stress detection compared to other multimodal sensor data. The performance difference between task-aligned and a 5-minute window was minimal, while longer windows significantly degraded performance, as they often spanned across multiple calls.
- We build and analyze a month-long, in-the-wild call-level multimodal dataset and a unified modeling pipeline that systematically compares windowing (task-aligned vs. fixed), modality composition (source-wise ablations; with/without self-reports), and personalization. We further quantify when personalization overtakes a general model: in time-ordered within-person splits, early folds underperformed or matched the general model, while later folds surpassed it.
- We explain and qualify the modeling patterns with interviews showing why CI and nCI aligned feature extraction helps, why per-person repertoires yield heterogeneous feature sets, and what lies beyond sensing coverage. From these, we discuss implications for when to intervene, how to personalize, and *how* to safeguard against surveillance in deployment.

2 Background and related works

2.1 Stress Sensing Using Sensor Data

In the HCI community, research has been conducted using multimodal data in various settings (e.g., laboratories, daily life, and driving) to detect stress and enhance individual well-being [27, 41, 45, 48, 78, 96]. Traditionally, stress detection involves collecting behavioral and physiological responses during stress-inducing tasks (e.g., the Trier Social Stress Test) in laboratory settings to build models for detecting stress states [86]. Stress assessments in uncontrolled, real-world environments are performed using physiological responses [53] or self-reported perceived stress [50]. Physiological stress responses, such as heart rate (HR), are considered an objective method for defining stress states [53]; however, they only measure immediate responses and fail to capture the cognitive stress states that persist after the stress stimuli are removed [11]. In contrast, commonly used self-reported stress levels, in which participants answer Likert scale questions about perceived stress, have been considered the most conventional label data for capturing the residual emotional and cognitive stress states after an immediate physical response [41, 50]. To detect perceived stress states, studies have explored data including computer interaction data (e.g., keyboard typing [38]), physiological data (e.g., HR [64], skin temperature [52], and skin conductance [52, 84]), behavioral data (e.g., accelerometer [52, 84] and mobile usage [84]), and context information that can influence stress (e.g., physical activity, location, movement, and environmental data [87]). These data were

used to develop classification models for automatically detecting stress in real environments [48].

2.2 Stress Sensing at the Workplace in HCI

The workplace is a prominent context in which research on stress sensing and related intervention systems is actively conducted, because workplace stress is recognized as a significant problem threatening individual and organizational well-being and incurring costs that must be addressed [28, 42]. Studies on sensing workplace stress typically utilize mobile devices and wearable devices, such as arm bands [88], wristbands [83, 90], and chest bands [67], to collect user data. These studies also integrate factors that reflect the workplace environment, such as primary work behaviors (e.g., email usage or engagement in productivity tools), into their data collection efforts [44, 59]. Also, they underscore the importance of incorporating workplace-specific data and considering the broader socio-technical context to ensure that such systems are effective [8, 21, 49].

However, existing sensor data research on workplace stress detection or data sharing has predominantly focused on the limited context of knowledge workers, such as information workers and employees of high-tech companies [11, 59]. Therefore, both the types of data collected and the collection protocols have been tailored to the characteristics of their work environment [59, 66]. Type of computer-interaction logs, for example, are limited to their task boundaries, and self-reported stress levels have typically been gathered at random intervals [88]. Although the accelerating digitalization of industry is gradually extending stress detection research to other high-stress risk groups (e.g., nurses [29], residents [97], construction workers [44]), some worker populations under high stress, such as call agents, have been explored only to a limited extent [37]. Therefore, to expand the benefits of data-driven research to such underexplored workforces (i.e., call agents), a study that reflects the nature of their work is required.

2.3 Emotional Labor and Stress

Emotional labor is the process of perceiving one's own emotions and managing emotional expressions in the workplace [39]. Various theoretical studies have been conducted in psychology to understand the mechanisms underlying emotional labor [7, 32]. Grandey described emotional labor through the 'emotional regulation' framework, emphasizing the choice of intrinsic strategies and the combined impact of situational cues, individual factors, and organizational factors [32]. Emotional labor is not limited to specific jobs [25]; however, it is a central task for individuals in customer-facing roles, such as flight attendants and call agents, collectively referred to as 'emotion workers' [40]. While sensor-based automatic stress detection research in this population remains limited, some studies have demonstrated the value of self-tracking emotional status for emotion workers [80]. Stress-sensing tools can contribute to the development of adaptive interventions, such as emotion regulation tools [89]. Nevertheless, existing studies fail to comprehensively address the multifaceted nature of emotional labor and the diverse stressors emotion workers face. For example, Hernandez et al. developed a model to classify stress conditions using skin conductance data from call agents [37], but their approach overlooked critical contextual, environmental, and organizational factors influencing stress. Park et al. tried to build a model to assess the intensity of emotional labor in a laboratory-simulated call center [73]. However, their findings do not fully account for the repetitive and enduring nature of emotional labor in real-world settings. Given these gaps, we examine call centers as one typical setting of front-line emotional labor where CI and nCI episodes and task logs are explicit. We now detail the study design and methodology.

3 Study Design and Methodology

To address our research question, we conducted a one-month, in-the-wild field study at a call center using a mixed-methods design approved by our Institutional Review Board (KH2022-108). This section details (1) the study context and participants, (2) quantitative data collection apparatus and procedures, (3) the interview protocol, and (4) the mixed-methods analysis pipeline.

3.1 Participants and Study Context

18 out of 24 call agents (female: 17, male: 1) participated in data collection for one month from June 22 to July 21, 2023, at the call center of a central metropolitan city government located in South Korea. The sex ratio of the participants is skewed compared to the general population; however, it reflects the reality of call center work, which is typically a female-dominated profession, as evidenced in the literature [10, 18]. It also closely mirrors a typical ratio in Korea, where male workers are typically less than 10% in call centers. The participants averaged 36.83 years (SD = 5.91) with 6.17 years of work experience as call agents (SD = 4.41). The city hall has 2–24 call agents working 8-hour shifts from 8:00 a.m. to 9:00 p.m. during weekdays and 3–4 agents working from 9:00 a.m. to 6:00 p.m. on weekends. Call topics include requesting information (e.g., transportation), complaints about city hall public affairs, and issues related to administrative tasks.

3.2 Data Collection Apparatus and Procedure

The detailed considerations of the data collection process, including ethical considerations, are provided in the Appendix A.1.

3.2.1 Sensors. Table 1 summarizes the sensors and their data types, and Figure 1 shows their placement.

Call log server. Modern call centers employ computerized call-management systems that automatically route calls and log detailed interaction data on company servers [98]. The system routes calls automatically using algorithms tailored to each company's policy, recording call information. Call agents use it to find information about callers, such as their call histories. These call log data provide rich contextual information about customer interactions and agent activities, which is crucial for analyzing stress levels. With company approval, we downloaded a de-identified subset, and call center managers manually verified that all personally identifiable information had been removed under IRB oversight. Each record captures start time, duration, inquiry type (e.g., complaint), the

Table 1: Summary of sensors

Sensor	Device	Frequency	Sensing Type	Data type	Data
Call log server	Servers managed by companies	Per call	Passive sensing	Task	Call start time, Call duration, Inquiry (text), Answer (text), Agreement, Complaint
Tablet	Galaxy Tab (A7 Lite 8.7)	Per call	Self-reporting	Behavioral	Eating, Drinking
		Per day	Self-reporting	Daily baseline	Daily health condition, Stress, Arousal, Valence, Bedtime, Wake–up time
		5.0 Hz	Passive sensing	Behavioral	Desktop activity (x,y, and z)
Environmental sensor	bluSensor (BSP02AIR)	0.1 Hz	Passive sensing	Environmental	CO_2, Humidity, Temperature
Wrist band	Fitbit Inspire 2	1 min	Passive sensing	Behavioral	Step counts
		1 sec	Passive sensing	Physiological	Heart rates

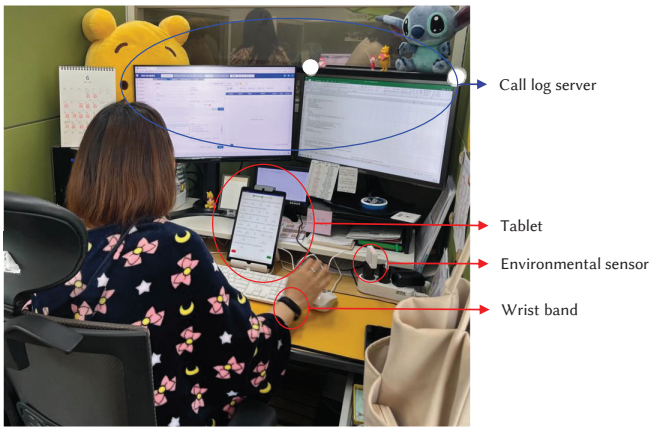


Figure 1: Data collection setup

agent's real-time notes, and—when customers consent via interactive voice response—an audio file.[1]

Fitbit. Fitbit is widely used to gather daily behavioral and physiological signals for stress detection [83]. Due to their compact display and extended battery life, they impose minimal participant burden during month-long in-situ deployments. Also, it provides a comprehensive set of preprocessed metrics. A limitation, however, is that heart rate variability (HRV), a pivotal biomarker of stress, is collected only during sleep. Although alternative wearables such as the Empatica E4 provide continuous HRV measurements, the present study employed the Fitbit Inspire 2, prioritizing ease of use for participants unfamiliar with smartwatches. We retrieved minute-level step counts and second-level heart-rate (HR) data via the Fitbit API; the former provided a baseline of physical activity before customer calls, while the latter captured stress responses during high-pressure interactions.

Environmental sensor. Environmental factors (e.g., high noise levels and poor air quality [26, 35]) are well-established workplace stressors [93]. Noise is especially relevant for call agents; however, the company's regulations prohibited the deployment of external ambient-noise sensors, so noise levels could not be recorded in our data collection. We instead monitored indoor CO_2, which has been shown to heighten physiological stress responses [92] and is also linked to verbal communication, because it is acceptable to the organization. bluSensor (BSP02AIR) was installed on every agent's desk to capture CO_2 along with other environmental conditions (e.g., temperature and humidity) at 10-second intervals and stream the data to a tablet. Although call agents share a single room, cubicle partitions can create micro-environmental differences; therefore, per-desk sensing improves spatial precision. Temperature and humidity were retained as complementary features to improve stress-detection models [95].

Tablet. Galaxy Tab A7 Lite (8.7 in.) tablets were placed next to each agent's keyboard to record desk activity via the built-in tri-axis accelerometer (ACC) at a rate of 5 Hz. Call agents spend most of their working hours at their assigned desks. Note that typing to retrieve information and log call outcomes is the most consistent and relevant behavior at the assigned seat. The ACC signal chiefly captures typing dynamics as an established stress marker [38], while still reflecting other micro-movements, such as reaching for documents or posture shifts. The same tablets were also administered two self-report protocols.

(1) Daily survey (before work). Call agents rated their baseline mental state, namely general health, arousal, valence, and stress, on 5-point Likert scales, as well as prior-night sleep quality (bedtime and wake-up time). All variables are known to modulate daily stress [63, 84].
(2) Post-task survey. After every call, they noted food or beverage intake since the last report. They can be behavioral responses under stress [77], and influential factors to the physiological responses (e.g., HR [86]). We asked participants to rate arousal, valence [86], surface acting (emotional labor strategy to display the required emotion externally), and deep acting (emotional labor strategy to change internal feelings into required emotion) [32] on a 5-point Likert scale.

[1]Audio was excluded from the present analysis because of its limited availability (related to callers' consent rejection).

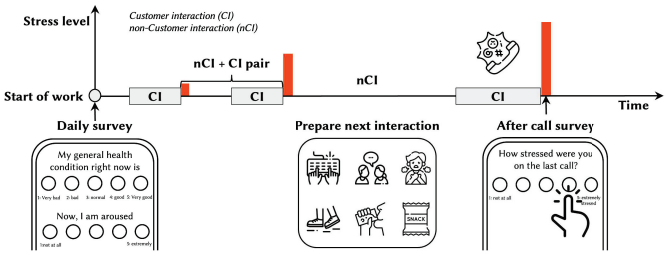


Figure 2: Daily workflow during data collection: Participants reported daily baseline before they start work. After each call (CI), they recorded their perceived stress levels before next call (nCI). During the nCI, participants engaged in various coping behaviors to prepare the next call voluntarily.

Note that these four post-task ratings were used solely to verify label reliability and were not fed into the stress-detection models.

3.2.2 Perceived Stress after Call. Self-reported stress measurements are a widely used method for acquiring label data in stress sensing research [85]. To collect ground truth, the participants were instructed to report their perceived stress levels on a 5-point Likert scale *after each call* ("How stressed were you on the last call?"). It directly addresses the stress from the most recent customer interaction, providing immediate and relevant information on their emotional states. Task-level self-reporting only after each call aligns well with the nature of call center work. Prior HCI studies also identified that such natural task breakpoints present optimal opportunities for interruption [2]. Unlike previous studies that often avoided frequent data collection to minimize participant burden [88], acquiring self-reported data after each call did not increase the workload of call agents, as we corroborated in the pilot study. Furthermore, given that each customer interaction is a separate event, this frequent reporting aligns naturally with the context of this study and enables us to capture the nuanced differences in stress responses following each interaction.

3.2.3 Procedure. Before initiating the main data collection, we sought feedback from organizations and participants during the recruitment stage to address potential legal restrictions and ethical concerns, such as excessive data collection that might compromise their privacy. Based on this feedback, we limited the data sources to devices agreed upon for use and data collected only during work hours. We conducted a one-day pilot study with five call agents, confirming that the procedure had minimal impact on their work. After confirming that the designated data collection method would not influence workers' work, researchers individually instructed participants on device usage, obtained consent forms, and collected demographic information through a pre-survey. Researchers positioned the sensors uniformly on each desk to maintain consistency among participants and asked them not to move the devices during the data collection period. During the data collection, participants were required to wear Fitbits during work hours and complete daily and after-call surveys using a tablet application. Figure 2 shows a daily workflow during the study. Referring to an existing incentive framework for data collection [61, 79], we required participants to meet minimum data collection standards for financial rewards (500,000 won, $358), along with additional incentives (100,000 won, $72) for the top two performers. Remote monitoring and systematic communication protocols were implemented to ensure data quality in the absence of on-site researchers. After data collection concluded, participants were voluntarily involved in an interview.

3.3 Interviews

We conducted semi-structured interviews with 13 participants to gain a deeper understanding of the collected data and its meaning in stress detection after data collection. Each interview was conducted individually in a separate space and lasted approximately one hour, with a reward of 10,000 KRW ($7.2). The manager arranged a schedule to ensure minimal disruption to their work responsibilities. In the interview, we asked the causes and responses to stress during work through the following questions:

(1) What triggers stress during CI?
(2) How do you respond to stressful conversations during CI?
(3) How do you prepare for the next call during nCI when a previous CI causes stress?
(4) Besides CI, which factor influences your perceived stress level after a call?

3.4 Data Analysis: Mixed-Methods Approach

We adopted a mixed-methods strategy to answer our research question. Quantitatively, we constructed a unified ML pipeline that standardizes preprocessing, feature engineering, and evaluation to enable comparisons across feature extraction windowing strategies, modality composition, and the effect of personalization. Qualitatively, we conducted an inductive content analysis of interviews not only to contextualize behavioral indicators from ML models in everyday work practices but also to surface mechanisms that sensors cannot capture. Here, we explain our quantitative pipeline and qualitative analysis protocol.

3.4.1 Quantitative Analysis: Preprocessing and Feature Engineering. We first processed all raw streams through a unified pipeline (Figure 3). The pipeline standardizes how raw data are cleaned, transformed into features, aligned to labels, and integrated into analysis-ready datasets, enabling comparisons across later modeling conditions. Additional implementation details are reported in Appendix A.2.

Data cleaning. We removed entries outside official work hours and screened the dataset for reliability issues at the participant and record levels. After an initial integrity check, we retained only participants and days with dependable self-reports and functioning devices. We then performed per-modality quality control: call logs were de-identified and de-duplicated under IRB oversight; heart-rate (HR) streams were checked for non-wear gaps and kept as missing when sparse; minute-level step outliers were screened using an isolation-forest procedure; tablet accelerometer (ACC) traces with device anomalies were excluded, and the remaining series were winsorized per participant using median absolute deviation; and environmental streams were checked for plausibility and obvious zeros.

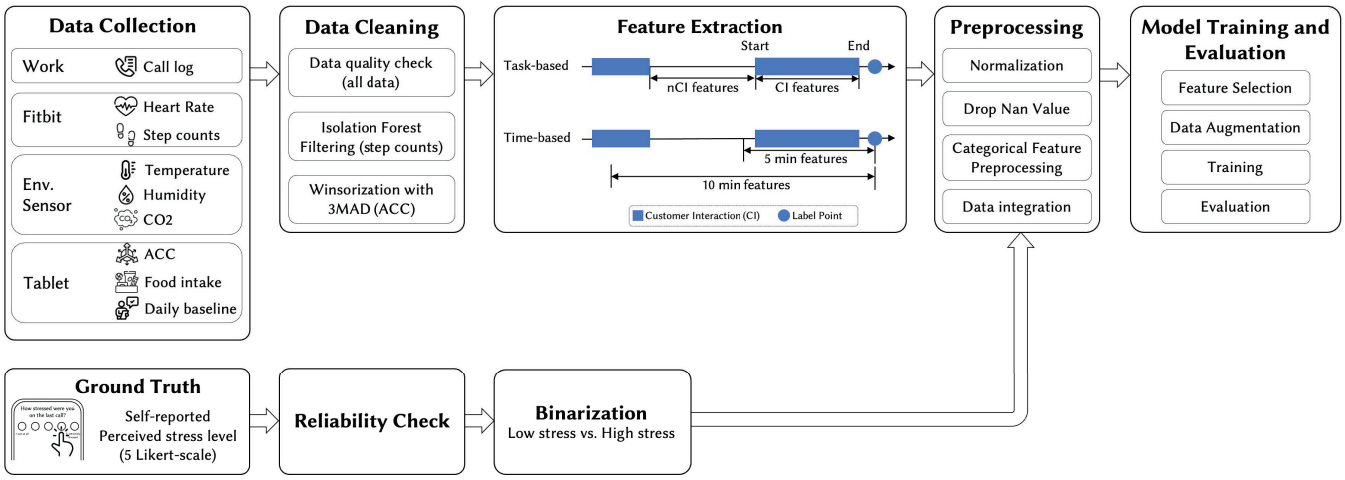


Figure 3: Data analysis pipeline. The pipeline includes data collection from multiple sensors, such as environmental (Env.) sensors and self-reported stress levels, followed by data cleaning, feature extraction using time-based and task-based (i.e., task-aligned) windows, and preprocessing. The dataset is split into subsets for feature selection and model training. Selected features are used for machine learning modeling and evaluation. Data augmentation is applied to address class imbalance.

Table 2: Features according to type of window

Sensors	Task-Aligned Window	Time-Based Window
Call log server	CI duration, nCI duration	Sum, mean, SD of CI and nCI duration
	Inquiry length, Answer length	Sum, mean, SD of inquiry/answer length
	Mute*	Total number of using mute
	Agreement type*, Complaint type*	Total number by agreement and complaint type
	Weekday*, Hour of day*	Sum of duration by agreement and complaint type
		Weekday*, Hour of day*
Tablet	Eating*, Drinking*	Total number of eating and drinking
	Daily health condition, Stress, Arousal, Valence, Sleep time	(same as left)
	Mean, SD, absolute integral of x, y, z in CI and nCI	Mean, SD, absolute integral of x, y, z, and magnitude
	Magnitude (CI and nCI)	—
Environmental sensor	Mean, SD of CO_2, humidity, temperature (CI and nCI)	Mean, SD of CO_2, humidity, temperature
Wristband	Mean, SD, sum of step count in nCI	Mean, SD, sum of step count
	Mean, SD of heart rate (CI and nCI)	Mean, SD of heart rate

Note: The categorical features are marked with *. Refer to Table 7 in Appendix A.2.2 for additional information.

Feature extraction. We implemented two distinct windowing strategies for feature extraction: task- and time-based windows (Table 2). *(1) task-aligned Window*: It aligns with the episodic nature of customer interactions (CIs) and non-customer interactions (nCIs) using the natural boundaries of each call or event. This method enables the extraction of context-specific features that reflect the inherent variability in the duration and activities of interaction [20]. *(2) Time-based Window*: It involves fixed-size windows that capture data from periods of 5 to 30 minutes, every 5 minutes before a label time point. While this method is conventional and widely used in studies where the temporal consistency of data collection is crucial [102], it also incorporates CI and nCI information from the calls within a given time window. This approach may only partly capture the episodic nature of emotional labor, as fixed window sizes may not fully align with the task-specific demands of CI and nCI. Nonetheless, it remains effective in providing a broader temporal context for feature extraction, especially when continuous monitoring is essential. We used the time information from the call log data as reference points for feature extraction. Definitions of each feature are in Appendix A.2.2.

Label processing. After-call perceived stress (5-point Likert) was matched to the immediately preceding call by UTC-timestamp. Labels recorded after the start of a subsequent call were discarded. To filter low-engagement responses, we applied a consistency screen [64] using the four post-call items (i.e., arousal, valence, surface acting, deep acting). To mitigate interpersonal differences in the baseline

of perceived stress level [37], we binarized stress per participant by thresholding at each person's mean [50]. We additionally verified robustness to alternative thresholding strategies (median-, and quantile-based), observing comparable performance (Appendix A.3).

Integration and one-hot encoding. For each windowing strategy, we joined features across modalities to their matched labels, removed rows with unresolved missing values, excluded all first-day data for participants to allow the experiment time to adjust to the setup and response methodology [51], and produced analysis tables (one task-based set; multiple time-based sets at different window lengths).

3.4.2 Quantitative Analysis: Modeling & Evaluation Design.

Nested LOSO Cross-Validation. We used nested leave-one-subject-out (LOSO) cross-validation to prevent test data leakage and optimistic bias. Within the training portion of each outer fold, we performed an inner LOSO loop that conducted scaling fits (i.e., normalization), feature selection, and hyperparameter tuning for models. The configuration achieving the highest mean inner-LOSO ROC-AUC was then retrained on the full outer-training set and evaluated once on the held-out participant. This procedure ensures strict separation between model selection and final testing.

Normalization. Numeric features were z-normalized using mean and standard deviation estimated from the training data within each CV fold, and the resulting transformation was applied to the held-out participant for test (i.e., the held-out participant's data were not used in testing).

Feature selection and data augmentation. LASSO [71] feature selection was applied only to models that are sensitive to multicollinearity. Extra feature selection was not performed for gradient-boosted decision tree models, which inherently perform feature selection and regularization as part of their hyperparameter optimization during training. Similarly, DL models were not manually restricted, as this could limit the model's ability to learn meaningful representations from heterogeneous modalities. To address class imbalance without evaluation bias, we applied SMOTE [15]/SMOTENC only within training folds and per participant to preserve within person structure. Feature selection and oversampling were fit on only training folds to prevent leakage.

Classifiers. Because our features are heterogeneous and tabular, we compared a compact set of ML classifiers spanning linear, kernel baselines and tree-based ensembles such as Decision Tree (DT) [54], Linear Discriminant Analysis (LDA) [9], Support Vector Machine (SVM) [36], Random Forest (RF) [12], eXtreme Gradient Boosting (XGBoost) [16] and CatBoost [76] as well as table-oriented DL models such as Tabnet [6], Tabtransformer [43], and NODE [75]. Detailed model setup (e.g., hyperparameters) appears in the Appendix A.2.6.

Training and evaluation. We used AUC-ROC, PR-AUC, and weighted F1 as primary metrics and performed non-parametric paired tests (i.e., Wilcoxon signed-rank test, $\alpha = 0.05$) for model performance contrasts. For stability, each evaluation was repeated over 30 times with fixed random seeds; means and SD per test folds are reported.

Comparative plan. Guided by our RQ, we designed a controlled comparison within the unified pipeline Leave-one-subject-out (LOSO) cross-validation.

(1) Windowing strategies for feature extraction: This aims to test the baseline performance of the collected dataset for comparison, determine whether task-aligned approaches better capture after-call stress than a fixed time window, and examine how time-window variation affects the model within a fixed-time window. To answer these questions, we evaluated all nine models across predefined feature sets as listed in Table 2 using the unified basic pipeline and analyzed the model performance. The top-performing model based on the AUC-ROC score is used for subsequent comparisons, so that differences reflect the representation (i.e., windowing, modality, and personalization) rather than model choice.
(2) Data source composition: This aims to identify which data type (i.e., task, behavioral, environmental, and physiological data) contributes most to after-call stress detection and evaluate the impact of sensing type (i.e., passive sensing vs. self-reporting). First, we conduct sensor ablation studies using only passive sensing data. Starting with a baseline configuration containing only task-related functions (i.e., call logs), we incrementally added environmental, behavioral, and physiological sensor data to evaluate model performance comparatively. Then, we assessed the effect of incorporating two types of self-reported data (i.e., after-call behavioral and daily baseline data) into each passive sensing configuration.
(3) Personalization: This study aims to investigate the performance of a personalized modeling approach under unified work conditions and to identify factors contributing to the differences in after-call stress detection. We train three types of personalized models with different train and test dataset partitioning strategies with time-based 5-fold cross-validation. First, we used only personal data that maintains chronological order, progressively increasing the training set. Second, we combined data from all participants into a single time-ordered dataset, disregarding participant boundaries and splits, to utilize the tested user's past data for training. Lastly, we used each participant's data equally partitioned across folds while preserving chronological order within each participant, thereby accounting for data balance across individuals. Additionally, we investigate the differences and commonalities of feature importance in modeling between the generalized model and the personalized model using the Shapley value from SHapley Additive ex-Planations (SHAP) [57].

3.4.3 Qualitative Content Analysis.

To provide deeper insight into the quantitative findings and find challenges beyond data, we conducted inductive content analysis [17] on the interview data (Section 3.3). The process unfolded in two stages of open coding and collaborative theme development. Initially, the lead analyst inductively coded transcripts and memos. This step involved reviewing the data line-by-line, identifying initial codes directly from the data. Particular attention was paid to how CI and nCI transitions shape the experience of call agents and their stressors and stress responses. Following the open coding, an additional researcher was involved

in reviewing the codes and engaging in collaborative theme development. Regular meetings were held between the two researchers to discuss the codes and elevate them into broader themes (e.g., prolonged CI as a stressor, personal difference under stressful CI; preparatory nCI routines). All interviews were audio-recorded and transcribed; Korean transcripts were coded in the original language, and translated excerpts are reported verbatim for quotes.

4 Results

We answer our research question by combining modeling evidence with interview findings. Following post-collection quality checks (Appendix A.2), we retained 15 from 18 initial participants for analysis: two agents exhibited near-constant low-stress self-reports (≥ 97% identical labels over 900 entries) and one self-reported temporal inconsistency in her labels. In addition, we also removed records outside official working hours. As a result, 7,442 call-level instances from 15 agents (310 days in total; 67.96% low stress, 32.04% high stress) using a unified pipeline. Results are organized along the planned quantitative comparison: (1) *windowing* (task-aligned vs. fixed), (2) *data-source composition* (source-wise ablations and the effect of adding self-reports), and (3) *personalization*. We then present qualitative themes that explain *why* the models behave and highlight considerable insights that our data may limit in capturing.

4.1 Performance Comparisons of Task-Aligned and Time-based Windowing

We applied the unified modeling pipeline to evaluate model performance across nine classifiers and compared feature extraction strategies (i.e., time-based and task-aligned, Table 3). Tree-based ensembles (i.e., RF, XGBoost) consistently outperformed other ML models and tabular-oriented DL models. Notably, the Random Forest (RF) model achieved the highest performance with an ROC-AUC of 0.692 and PR-AUC of 0.518 in the 5-minute window setting. Given the class imbalance, these PR-AUC scores significantly exceed the random baseline of 0.32. Across fixed time windows, a clear inverse trend was observed; performance declined as the window length increased. For instance, the ROC-AUC of RF dropped from 0.692 (5-min) to 0.605 (30-min), suggesting that shorter windows capture transient stress markers more effectively. Meanwhile, the 5-minute window yielded performance comparable to the task-aligned strategy, which aligns well with the fact that the average call duration is shorter than 5 minutes. Statistical analysis (Wilcoxon signed-rank; Appendix A.4) confirmed that these two top-performing configurations are statistically indistinguishable across all metrics. PR-curve in Figure 4 visualizes the model performance to classify the high stress condition in precision and recall space: the task-aligned curve closely overlaps with the 5-min window across thresholds, whereas longer windows shift downward, indicating lower precision at the same recall under class imbalance.

4.2 Which Data Sources Matter?

We conducted a data source ablation study to identify the most informative data sources and assess the added value of self-reported features in model performance. We decided to use one of the best-performing configurations from Section 4.1 (i.e., RF with task-aligned features). Since the task log features are naturally aligned

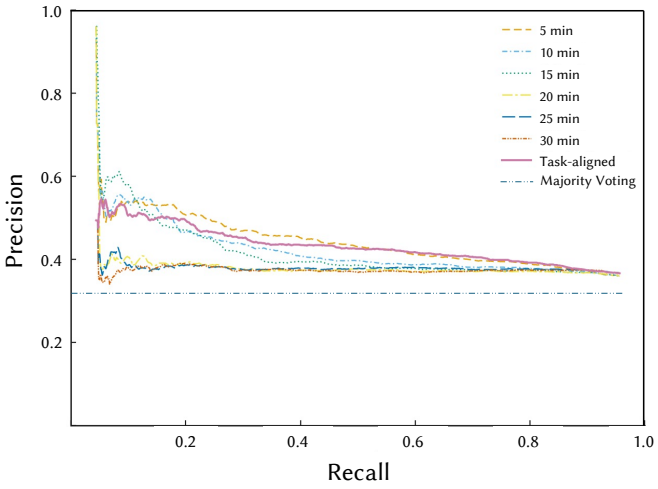


Figure 4: PR-AUC curve of each windowing strategy.

with the task cycles (CI and nCI), task-aligned windowing was a more natural choice for evaluating the impact of data sources. We evaluated a total of 15 feature set combinations spanning task logs, tablet ACC, HR/steps, environmental streams (CO_2, temperature, and humidity), and optional self-reports (after-call behavioral item, and daily baseline). The gap between the best configuration (i.e., all sources) and the worst (i.e., excluding task logs) was ΔROC-AUC = 0.038 (6.15% relative increase, Figure 5; Appendix A.5). Removing task log features resulted in a substantial performance decrease (ΔPR-AUC = 0.047). As shown in the Precision-Recall curves (Figure 6), the model including task features consistently maintains higher precision across the entire recall range compared to the no-task baseline, confirming the robustness of task signals. The addition of self-reported data had negligible effects across most passive sensing combinations. These patterns align with interview accounts that stressful customer interactions are prolonged and complex, thus surfacing in *task logs* such as call inquiry length (Section 4.4.2). Note that similar patterns were observed in other models with fixed windowing, such as 5-minute windowing; we omit the results for the sake of brevity.

4.3 Personalization: Interpersonal Variability

Across the three personalization schemes, performance generally improved as more personal history added (Figure 7a, 7b).[2] Early folds (Folds 1–3) underperformed or matched the leave-one-subject-out (LOSO) baseline, whereas later folds outperformed the general model performance in general. In particular, the hybrid variants exhibited a pronounced improvement from Fold 4 onward in both ROC-AUC and PR-AUC (Figure 7), while the purely personalized model improved up to Fold 4 but showed larger fold-to-fold variability.

To quantify how much personal history is needed before personalization tends to outperform the general (LOSO) model, we define the *crossover point* as the earliest time-ordered fold at which

[2]Each test was repeated 30 times with fixed random seeds. Error bars (when present) denote 95% confidence intervals.

Table 3: Model performance by windowing strategies (LOSO)

	Task-aligned	Time-based					
		5-min	10-min	15-min	20-min	25-min	30-min
ROC-AUC mean (STD)							
RF	**0.685** (0.090)	**0.692** (0.082)	0.671 (0.080)	0.642 (0.074)	0.616 (0.074)	0.610 (0.067)	0.605 (0.0638)
XGBoost	0.679 (0.085)	0.686 (0.088)	0.664 (0.083)	0.637 (0.082)	0.625 (0.072)	0.613 (0.078)	0.599 (0.057)
CatBoost	0.661 (0.094)	0.673 (0.085)	0.656 (0.085)	0.627 (0.084)	0.600 (0.086)	0.591 (0.075)	0.594 (0.072)
LDA	0.661 (0.079)	0.663 (0.081)	0.643 (0.078)	0.617 (0.071)	0.599 (0.069)	0.593 (0.066)	0.589 (0.065)
SVM	0.623 (0.104)	0.624 (0.106)	0.612 (0.099)	0.584 (0.099)	0.567 (0.093)	0.560 (0.094)	0.546 (0.092)
DT	0.576 (0.071)	0.601 (0.066)	0.577 (0.068)	0.583 (0.047)	0.536 (0.031)	0.523 (0.039)	0.547 (0.044)
TabNet	0.658 (0.039)	0.676 (0.035)	0.636 (0.030)	0.612 (0.048)	0.587 (0.042)	0.595 (0.025)	0.587 (0.015)
TabTransformer	0.644 (0.038)	0.664 (0.034)	0.645 (0.038)	0.624 (0.039)	0.603 (0.035)	0.609 (0.038)	0.601 (0.035)
NODE	0.640 (0.050)	0.646 (0.040)	0.536 (0.0450)	0.607 (0.049)	0.582 (0.040)	0.586 (0.039)	0.566 (0.034)
Majority Voting	0.500						
PR-AUC mean (STD)							
RF	0.514 (0.197)	0.518 (0.197)	0.510 (0.190)	0.475 (0.181)	0.433 (0.175)	0.423 (0.164)	0.417 (0.162)
XGBoost	0.516 (0.192)	0.525 (0.192)	0.508 (0.182)	0.477 (0.169)	0.452 (0.172)	0.426 (0.170)	0.421 (0.162)
CatBoost	0.491 (0.191)	0.510 (0.197)	0.494 (0.193)	0.473 (0.182)	0.437 (0.183)	0.423 (0.176)	0.422 (0.172)
LDA	0.514 (0.179)	0.517 (0.180)	0.496 (0.166)	0.465 (0.158)	0.443 (0.163)	0.439 (0.154)	0.432 (0.152)
SVM	0.481 (0.201)	0.475 (0.203)	0.469 (0.198)	0.443 (0.185)	0.418 (0.178)	0.414 (0.172)	0.405 (0.172)
DT	0.399 (0.159)	0.411 (0.170)	0.404 (0.159)	0.394 (0.162)	0.356 (0.151)	0.348 (0.141)	0.368 (0.154)
TabNet	0.486 (0.103)	0.512 (0.096)	0.476 (0.086)	0.449 (0.088)	0.422 (0.092)	0.420 (0.085)	0.412 (0.085)
TabTransformer	0.490 (0.092)	0.517 (0.091)	0.502 (0.089)	0.479 (0.082)	0.443 (0.086)	0.453 (0.078)	0.441 (0.075)
NODE	0.478 (0.099)	0.476 (0.476)	0.643 (0.093)	0.434 (0.090)	0.409 (0.088)	0.416 (0.085)	0.402 (0.078)
Majority Voting	0.320						
Weighted F1 mean (STD)							
RF	0.617 (0.168)	0.631 (0.128)	0.608 (0.150)	0.597 (0.149)	0.579 (0.145)	0.574 (0.151)	0.577 (0.152)
XGBoost	0.601 (0.145)	0.610 (0.142)	0.601 (0.160)	0.571 (0.153)	0.578 (0.146)	0.570 (0.147)	0.567 (0.148)
CatBoost	0.596 (0.155)	0.609 (0.147)	0.590 (0.157)	0.574 (0.161)	0.568 (0.148)	0.557 (0.150)	0.553 (0.148)
LDA	0.597 (0.120)	0.586 (0.132)	0.571 (0.146)	0.540 (0.145)	0.531 (0.148)	0.522 (0.14803)	0.516 (0.146)
SVM	0.566 (0.162)	0.571 (0.154)	0.562 (0.154)	0.545 (0.153)	0.537 (0.144)	0.539 (0.148)	0.533 (0.143)
DT	0.579 (0.094)	0.588 (0.071)	0.588 (0.089)	0.592 (0.075)	0.560 (0.076)	0.551 (0.0841)	0.581 (0.0949)
TabNet	0.594 (0.046)	0.589 (0.035)	0.583 (0.038)	0.561 (0.062)	0.509 (0.045)	0.523 (0.049)	0.508 (0.034)
TabTransformer	0.582 (0.049)	0.575 (0.056)	0.530 (0.054)	0.512 (0.082)	0.535 (0.053)	0.508 (0.063)	0.548 (0.067)
NODE	0.584 (0.065)	0.605 (0.055)	0.537 (0.080)	0.543 (0.082)	0.526 (0.076)	0.527 (0.077	0.497 (0.078)
Majority Voting	0.550						

performance exceeds the LOSO baseline at the group level. This crossover occurs at Fold 4. The amount of personal history available for training corresponded to a mean of 329.6 calls per participant (median 332; range 92–612). At the participant level, 9 of 15 participants outperformed the LOSO baseline in ROC-AUC and 8 of 15 in PR-AUC at Fold 4, while 5 of 15 did not surpass the baseline on either metric, indicating substantial interpersonal variability. Notably, call volume alone was not strongly discriminative (median calls: 300 for participants surpassing the baseline on both ROC and PR vs. 284 for those surpassing neither), and some participants never exceeded the baseline even at their best fold. Practically, we therefore recommend deploying the general (or balanced-hybrid) model during onboarding and attempting personalization once roughly ∼300–330 personal calls are available, switching only when participant-specific validation indicates an improvement over the general baseline.

At the feature level, personalization surfaced heterogeneous sets rather than a single universal signature (Table 4). Across 15 participants, no feature appeared in every top 10. The composition in personalization spanned task logs (i.e., inquiry length) as well as environmental (i.e., 'SD of humidity', 'SD of temperature', and 'SD of CO_2'). This contrasts with the general model, which relied more heavily on task-log features and other data collected during CI (Table 5).

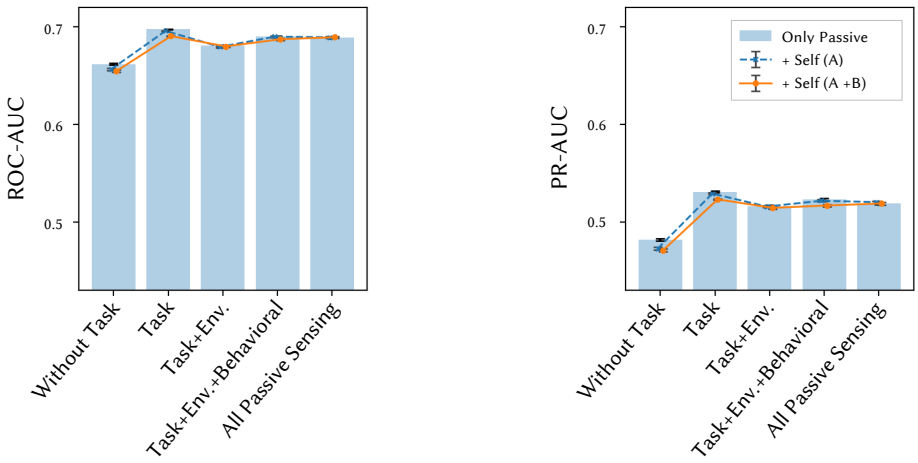


Figure 5: Ablation results comparing multiple combinations of passive sensing alone to combinations with self-reported data (A and A+B). Bar plots show ROC-AUC (left) and PR-AUC (right) across sensing subsets: Without Task, Task-only, Task+Environmental, Task+Environmental+Behavioral, and All Passive Sensing. Lines indicate performance changes when adding after-call self-reports (A) or adding both after-call and baseline self-reports (A+B).

Table 4: The important features in personalized models

Sensors	Features	Period	# of ranked
Tablet ACC	int of y	CI	6
Call log server	inquiry length	CI	6
bluSensor temperature	SD of temperature	CI	6
bluSensor humidity	SD of humidity	CI	6
Tablet ACC	int of z	CI	5
Tablet ACC	mean of magnitude	nCI	5
bluSensor temperature	SD of humidity	nCI	5
Tablet ACC	int of x	CI	4
Call log server	complain 1	CI	4
bluSensor CO_2	mean of CO_2	nCI	4

Table 5: The important features in general model

Sensors	Features	Period
Call log server	complain 1	CI
Tablet ACC	int of y	CI
Call log server	duration	CI
Call log server	agreement 1	CI
Tablet ACC	int of z	CI
Tablet ACC	int of x	CI
Tablet ACC	int of mean	nCI
Call log server	answer length	CI
Call log server	inquiry length	CI
Call log server	agreement 0	CI

4.4 Qualitative Themes: Why These Patterns Emerge and What Exists Beyond Our Sensing

These qualitative findings not only contextualize the quantitative results but also reveal points where model predictions diverge from workers' lived experience. In several cases, workers' accounts contradict what is captured in logs or sensor streams, highlighting conceptual blind spots in the modeling pipeline. It also highlights that self-reporting stress after calls provided emotional awareness, aiding preparation for next interactions. Below, we present themes that (1) *support* why task-aligned representations and task logs work well in our models, and (2) *challenge* what these signals can capture by surfacing sources of divergence such as residual stress during nCI, opposite behavioral signatures, and unobserved embodied or contextual stressors.

4.4.1 Task Segmentation: Dividing Customer and Non-Customer Interactions. Agents described nCI as the moment to reset display rules and prepare for the next CI. Many engaged in brief routines during nCI ($N = 11$)—for example, stepping away, briefly chatting with colleagues, taking deep breaths, or forceful typing. They also noted that the outcome of a previous call could affect the next one ($N = 8$), particularly noting the experience of insufficient time to ventilate stress from a previous call ($N = 4$): (P15) "*It was really tough when there was a short interval between calls.*" (P14) "*It's impossible to ventilate before every call.*" These accounts indicate clear mental and behavioral differences between CI and nCI that can be reflected in our data (e.g., desk activity and steps during nCI vs. call-centric logs during CI) and support using *task-based (episode-aligned) windows* that follow natural task boundaries (Section 4.1). However, these same accounts also caution against treating nCI as a uniform rest baseline: residual stress and active coping within nCI may persist into subsequent calls, contributing to temporal variability and potential divergence between sensed signals and post-call self-reports.

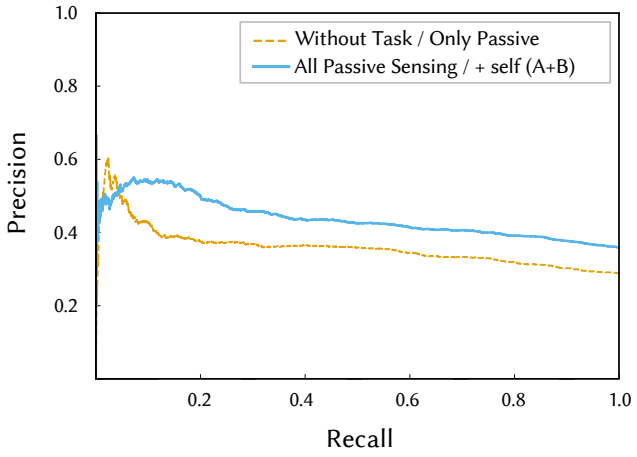


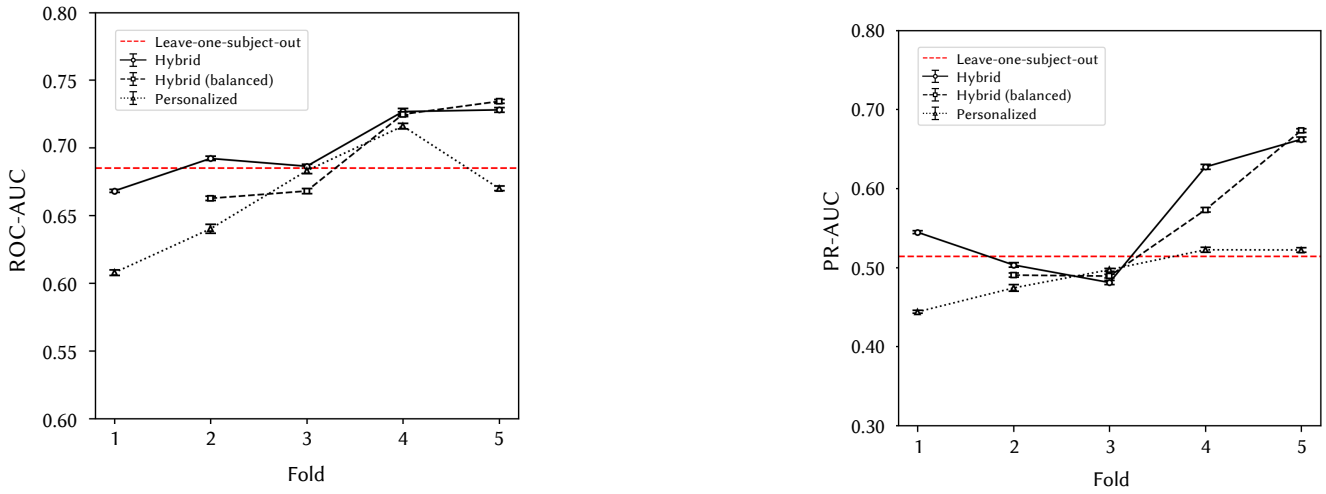
Figure 6: Precision–Recall curves comparing two ablation conditions: passive sensing without task features (yellow dashed line) versus all passive sensing combined with self-reported data (A+B; blue solid line). Adding task features and self-reports improves precision across nearly the entire recall range, while the passive-only model maintains lower and more rapidly declining precision.

4.4.2 How Stressful Customer Interactions Surface in Task logs. The participants reported that stress rose when customer problems remained unresolved or ambiguous ($N = 8$), when answers had to be repeated because callers did not understand or accept them ($N = 3$), and when conversations demanded more time for empathy beyond providing a solution ($N = 3$). The situations involved customers who did not know their problem or whose communication style failed to clarify the issue ($N = 3$) directly: (P4) "*Even though it's stressful when the customer is unpleasant, it's even more challenging when they inquire without knowing their issue. The process of figuring it out is complicated, and it becomes harder if they are not cooperative.*" It is also problematic when customers fail to understand or accept the proposed solutions ($N = 3$). (P10) "*I answered, but if they don't understand and I have to repeat the same content several times ... it becomes stressful.*" (P15) "*It's stressful when elderly callers keep repeating the same content over the phone.*" Even when the problem is clearly identified, customers express emotions unrelated to the solution or demand empathy, which leads to additional communication requirements ($N = 3$): (P8) "*After venting to me for about 11 minutes on a call ... They finally said, 'Talking about it like this has made me feel a bit better.' I pretended to be happy, but really feelings hidden* ***behind that smile*** *really hurt.*" These conditions surface in the task logs. Protracted explanation/empathy increases the length of agent-entered inquiry and response notes (i.e., inquiry length and answer length). Consistent with this mapping, call-centric features from task logs ranked highest in our general models, and removing task logs produced the largest AUC drop (Section 4.2). At the same time, call agents' accounts emphasize emotional dissonance (e.g., suppressed feelings) that may not scale monotonically with duration or note length, suggesting a concrete source of dissonance when models rely heavily on structural log proxies.

4.4.3 Individual Differences: Stillness vs. Fidgeting and Diverse Preparation. Several call agents described reduced movement during stressful interactions ($N = 7$), while others reported active coping ($N = 5$). Inactive response was linked to intense focus or a sense of powerlessness. Note that when they feel powerless, they give up actions to find solutions and choose to endure the situation (N = 5): (P1) "*I slump and zone out because there is nothing I can do in the middle of the call.*" (P10) "*I start to give up if it feels hopeless, even if I first try to type and resolve it quickly.*" Active coping during CI included sighing ($N = 2$), forceful gestures or swearing under the desk ($N = 1$), and drinking water ($N = 1$). Individual differences were also salient in the after-call interval (nCI), which agents framed as time to reset and prepare for the next CI. While waiting for the next call, some talked with colleagues around their seats immediately after a call to relieve stress ($N = 2$), others sat alone, taking deep breaths ($N = 2$), typing rapidly and forcefully ($N = 2$), or consuming sweet snacks or cold beverages ($N = 2$). Notably, many reported having to leave their offices ($N = 7$), describing this action as a response to emotions that were too difficult to manage or when stress was overwhelming (N = 4). (P7) "*Getting up from my seat usually means I'm really stressed, and I need to step out even before I log the content of the finished call.*" These person-specific behaviors imply opposite signatures in the same modality, which helps explain why personalized models surfaced different top features by participant and drew on a broader mix of sources than the general model (Section 4.3). This semantic ambiguity provides a qualitative explanation for why a single global mapping underperforms some individuals. The same modality such as capturing physical movements during CI can encode opposite stress responses across agents, producing predictable error modes without per-person calibration.

4.4.4 Beyond Customer Interaction: External Factors Shaping After-Call Stressors. Beyond interaction content, after-call stress was shaped by individual traits (e.g., active temperament, low tolerance for repetition; $N = 5$) and day-level conditions such as poor sleep or throat pain ($N = 5$). Some female agents reported greater sensitivity during their menstrual period ($N = 4$). Environmental noise was also mentioned, but the specific triggers differed ($N = 3$). For example, among those who noted auditory environmental impacts (N = 3), the reasons differed: one mentioned the loud voices of colleagues handling aggressive complaints, another cited chatter from nearby colleagues, and a third pointed to the echoing of their own consulting voice in an overly quiet environment, or sounds revealing their activities, such as opening cans. Some modifiers were partially represented in our streams (daily baseline self-reports; desk-level CO_2, temperature, and humidity), whereas others (e.g., menstrual cycle and fine-grained acoustic context) were not directly measured. These unmeasured or partially measured factors plausibly shift within-person baselines over time, offering a qualitative account of intrapersonal drift such as daily performance variability.

4.4.5 Benefits of Task-Oriented Stress Self-Reporting. While automated detection was our primary focus, participants also reported ancillary benefits of the call-aligned self-reports that we used as



(a) ROC-AUC Trends **(b) PR-AUC Trends**

Figure 7: Performance validation across 5 folds. The proposed Hybrid model (solid line) shows consistent improvement over baselines in both (a) ROC-AUC and (b) PR-AUC metrics.

labels ($N = 12$). Some participants reported that rating stress after each call increased their in-the-moment awareness ($N = 5$), and others noted that it helped them emotionally prepare for the next interaction ($N = 3$). Because the prompts coincided with natural task boundaries, agents did not perceive additional workload. These observations account usefulness of semi-automated self-tracking that blends automated sensing with lightweight self-reports [19]. Participants' descriptions also suggest that repeated call-aligned self-reporting can function as a mild intervention (increasing awareness and supporting reset), which may gradually shift rating criteria over time. Such label dynamics can contribute to apparent model label dissonance.

5 Discussion

5.1 Task Logs as Primary Signals and the Value of Task-Aligned Segmentation

Our findings position *task logs* as primary sensing signals for after-call stress in emotional labor. In SHAP summaries and ablations of the general models, features derived from the task logs themselves consistently dominated. Interviews explain that stressful customer interactions were often prolonged, involved repeated explanations, or demanded empathy, all of which extend calls and increase strain. Thus, features such as inquiry length capture the volume and specificity of customer requests; both operationalize situational cues that heighten emotional-labor demands [32, 33]. A further benefit of log-based signals is their *semantic legibility* to workers and practitioners. Features such as call duration, holds, or repeated explanations map onto recognizable work events, making model outputs easier to interpret, contest, and act on. This readability helps organizational managers (e.g., human resources) understand meaningful workflows (e.g., call-specific reflections and trigger identification), enabling the organization to implement meaningful responses to manage worker health across the organization [49]. This allows for more intuitive and negotiable inferences compared to complex data requiring sensitive and meaningful inferences.

Beyond the sensed signal, task-aligned features outperformed longer fixed windows and remained competitive with short ones. Emotional-labor theory accounts for this advantage: transitions into CI act as situational cues that trigger regulation demands, whereas nCI supports preparation and recovery [32]. As such, segmentation of CI and nCI is not a preprocessing detail but a modeling choice that better matches the lived experience of emotional labor, which is inherently episodic, while existing stress-sensing research often did not explicitly segment. This episode-based framing also offers a natural *unit of intervention* [65]. CI marks moments where support must be minimally disruptive, whereas nCI provides actionable breakpoints for recovery-facing micro-interventions and reflection. We note, however, that emotional strain can vary dynamically within an episode [30], motivating future work on finer-grained, within-call sensing when it unlocks qualitatively new support.

5.2 Sensing for Emotional Labor: Mapping Modalities to a Demand, Regulation, Recovery Cycle

Our findings allow us to revisit how to conceptualize sensing stress in the context of "emotional labor." Building on emotional-labor theory [32], we argue that emotional labor unfolds episodically through *interaction demand* (i.e., situational cues), *regulation effort* (i.e., acting strategies enacted during interaction), and *recovery* (i.e., between-episode restoration) in our CI/nCI segmentation (i.e., CI captures regulation under demand, whereas nCI captures preparation and recovery). This framing helps interpret why task logs perform so strongly. In our setting, after-call stress is tightly coupled to interaction demand that is already reflected in operational traces. Consistent with prior workplace HCI work treating work-specific digital traces as de facto sensors [42, 55, 72, 101], our results extend

this line to emotional labor by sensorizing call task logs. Moreover, it offers semantic legibility for sense-making such as mapping features (e.g., duration, holds, or repeated explanations) onto recognizable work events, supporting interpretability, contestation, and action [80].

At the same time, the added values of multimodal sensing should be argued in terms of *actionability and meaning*, which go beyond merely improving AUC. Multimodal signals can enable finer temporal resolution (e.g., within-call escalation) and measure aspects of *regulation and recovery* that logs cannot observe, which can be critical for personalization and recovery-aware support [61, 66, 73]. Consequently, we argue against a binary choice between logs and sensors in favor of *progressive sensing*. It starts with low-intrusion, log-based sensing to address demand-driven strain, and adds modalities selectively only when they unlock qualitatively new support (e.g., recovery-aware microbreaks) or improve worker-facing interpretability and agency. Because additional sensors also introduce friction, privacy risk, and function creep in quantified workplaces [3, 21, 49, 81], deployments should prioritize worker control over what is inferred and shared, and favor privacy-preserving, negotiated uses over managerial surveillance.

5.3 Individual, Team, and Organizational Loops from Sensing to Intervention

We situate these applications in workplace *digital wellbeing* research that supports self-regulation and reflection while preserving user agency [34, 42]. Because workplace sensing can also enable *worker monitoring*, we design these loops around data minimization, worker control, and purpose limitation (see Section 5.4) [3, 21, 81].

Individual loop (worker-facing). Predictions can support *just-in-time* (JIT) micro-interventions at natural breakpoints (e.g., immediately after demanding calls), such as paced breathing, brief grounding prompts, or short recovery routines [28, 34, 42, 65]. Beyond on-the-spot support, model outputs can function as personal informatics (PI), helping workers review stress trajectories and identify personal triggers over time [47]. Coupling predictions with explainable AI can further surface interpretable cues (e.g., prolonged duration, repeated explanations, voice changes) that contributed to high stress, supporting worker sense-making and exploration of coping strategies [46]. To mitigate imperfect model performance while keeping burden low, a semi-automated approach that blends passive sensing with lightweight, call-aligned self-reports can provide semantic calibration and improve acceptance [19, 56, 69].

Team/workflow loop (coordination-facing). In call-by-call workflows, stress sensing can inform team-level load balancing and buffering (e.g., inserting short recovery windows after high-demand calls, recommending optional cooldown periods, or suggesting voluntary routing preferences such as "no call assignment for 15 minutes"). Importantly, these coordination supports need not reveal a worker's raw inferred stress state; instead, they can operate via *negotiated, worker-controlled signals* (e.g., "needs a short break"), reducing surveillance risk while still enabling practical scheduling and routing adjustments.

Organizational loop (policy/resource-facing). Aggregated, privacy-preserving analytics can help identify systemic stressors (e.g., understaffed periods, call types that routinely drive strain, or training gaps) and motivate organizational interventions such as staffing adjustments, targeted training, or access to debrief resources. Crucially, these uses require governance safeguards and purpose limitation to prevent repurposing stress inference for performance evaluation or disciplinary monitoring (see Section 5.4). These organizational uses should prioritize system-level improvements over individual attribution, and avoid individual-level managerial dashboards.

5.4 Ethical Usage of Stress Detection at Workplaces

Although workplace stress detection systems offer significant employee benefits as reported in prior studies [11, 28, 42, 60, 66], it is essential to carefully consider the ethical implications of quantified workplaces, particularly regarding privacy and surveillance, as passive sensing technologies continue to advance [22]. As confirmed in our findings, workplace stress detection heavily relies on work activity data (e.g., call logs); its continuous collection may cause workers to feel constantly monitored and pressured to manage their emotions, even to the sensors [81], resulting in additional stress. Such tensions can undermine the primary goal of promoting individual well-being, as workers still worry about potential misuse by organizational stakeholders [21].

However, recent studies have proposed possible solutions, emphasizing the importance of human-centered design, careful consideration of worker impacts, and system tailoring to mitigate potential harm [21]. Building on these approaches, we argue that stress detection models can be ethically utilized in the workplace if they are carefully designed, and considerations for ethical usage should be made in the development of future workplace stress detection models. First, systems that support the operation of such models must be designed with clear definitions regarding the scope of data collection, stored data management policies, access controls, and the transparency and explainability of trained models. To ethically establish these definitions, a user-centered design approach that actively involves participants throughout the entire design process must be employed [91]. Second, strong institutional and legal standards for ethical usage (e.g., securing user consent and defining clear system goals) must be established and enforced. For example, organizational policies and applicable regulatory frameworks should restrict stress inference to wellbeing-supportive purposes and prohibit use for performance evaluation or disciplinary monitoring. Lastly, a practical safeguard is to *separate outputs by level*. Individual-level inferences should remain worker-facing (e.g., personal informatics and worker-controlled signals), while organizational insights should be limited to privacy-preserving aggregates that target demand-side fixes (e.g., staffing or training) rather than evaluating individuals. Such purpose limitation reduces incentives for surveillance and helps preserve worker autonomy [91]. Ultimately, developing stress-sensing systems with consideration for ethical implications can ensure the privacy and autonomy of workers, empowering them without introducing additional layers of workplace surveillance [91], thereby enhancing their mental well-being.

6 Limitations

While our study offers significant insights into building an after-call stress detection model for call agents by collecting real-world multimodal sensor data, several limitations exist regarding our sensing choices and contextual constraints. First, our sensing choices were heavily influenced by the operational realities of a real-world call center. As detailed in Appendix A.1 (Table 6), we encountered strict regulatory restrictions on data types and significant privacy concerns regarding customer consent. Consequently, we intentionally excluded audio, video recordings of calls, and personal mobile data usage, despite their potential richness for emotion analysis, to prioritize agent privacy and comply with institutional security policies. Future research could explore privacy-preserving audio analysis techniques or simulated environments where such constraints are minimized, although this may come at the cost of ecological validity. Second, our findings should be interpreted within the specific cultural and gendered context of the study site, a predominantly female call center environment [18] characterized by strict emotional display rules requiring the suppression of negative emotions [99]. Such high-intensity emotional labor and organizational norms likely influence physiological stress patterns differently than in other contexts. For instance, the strictness of supervisory enforcement of display rules is linked to specific patterns of emotional exhaustion [94], while cultural values can moderate the physiological costs of emotional suppression [14]. Therefore, as suggested by recent cross-cultural sensing research [62], further validation is required to determine the transferability of our findings to call centers with different demographic compositions and emotional labor expectations. Finally, relying on self-reported stress levels via Ecological Momentary Assessment (EMA) introduces potential response bias. Although we implemented rigorous filtering to exclude insincere responses (e.g., straight-lining), future work should adopt systematic procedures to address the biases inherent in self-reports [74] and consider incorporating expert ratings or operational metrics as complementary ground truth.

7 Conclusion

This study extends prior research on automated stress detection for knowledge workers to emotion workers by presenting a method for detecting stress using real-world data. Overall, this study underscores the significance of considering the distinct work environment of emotion workers, particularly in terms of their susceptibility to stress. Specifically, we discovered that task-related features from call logs as a new sensor modality are crucial for model performance, and customer interaction features play a significant role in stress detection in general. Task-aligned windowing proved effective, as setting longer windows often degraded performance by spanning multiple calls, whereas shorter windows provided more accurate, context-specific stress detection. Additionally, we observed significant individual differences in personalized models, enhancing the effectiveness of tailored interventions for emotion workers. While the current work has limited generalizability, future research should expand to explore a large-scale, multinational dataset with voice data for a more comprehensive analysis. Note that we advocate for deploying stress detection tools (e.g., just-in-time intervention, semi-automated self-tracking) that prioritize improving employee well-being without contributing to invasive surveillance practices. Ongoing research should refine these tools to ensure they are ethically sound and practical for real-world applications.

Acknowledgments

The corresponding author of this work is Uichin Lee. This research was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (RS-2022-II220064) and by the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) (RS-2022-NR068758). The co-last author of this work is Vedant Das Swain. He was supported through Microsoft's Accelerating Foundation Models Research (AFMR) and their call to AI, Cognition, and Economy (AICE) research network. He also received additional support from NIH National Institute of Drug Abuse under award number NIH/NIDA P30DA029926.

References

[1] 360iResearch. 2024. *Call Center Market by Component, Deployment, Vertical - Global Forecast 2025-2030.* Technical Report. Research and market.
[2] Piotr D Adamczyk and Brian P Bailey. 2004. If not now, when? The effects of interruption at different moments within task execution. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 271–278.
[3] Daniel A Adler, Emily Tseng, Khatiya C Moon, John Q Young, John M Kane, Emanuel Moss, David C Mohr, and Tanzeem Choudhury. 2022. Burnout and the quantified workplace: Tensions around personal sensing interventions for stress in resident physicians. *Proceedings of the ACM on Human-computer Interaction* 6, CSCW2 (2022), 1–48.
[4] Daniel A Adler, Vincent W-S Tseng, Gengmo Qi, Joseph Scarpa, Srijan Sen, and Tanzeem Choudhury. 2021. Identifying mobile sensing indicators of stress-resilience. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5, 2 (2021), 1–32.
[5] Douglas G Altman and Patrick Royston. 2006. The cost of dichotomising continuous variables. *Bmj* 332, 7549 (2006), 1080.
[6] Sercan Ö Arik and Tomas Pfister. 2021. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 6679–6687.
[7] Blake E Ashforth and Ronald H Humphrey. 1993. Emotional labor in service roles: The influence of identity. *Academy of management review* 18, 1 (1993), 88–115.
[8] Ezra Awumey, Sauvik Das, and Jodi Forlizzi. 2024. A Systematic Review of Biometric Monitoring in the Workplace: Analyzing Socio-technical Harms in Development, Deployment and Use. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 920–932.
[9] Suresh Balakrishnama and Aravind Ganapathiraju. 1998. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing* 18, 1998 (1998), 1–8.
[10] Vicki Belt, Ranald Richardson, and Juliet Webster. 2002. Women, social skill and interactive service work in telephone call centres. *New technology, work and employment* 17, 1 (2002), 20–34.
[11] Brandon M Booth, Hana Vrzakova, Stephen M Mattingly, Gonzalo J Martinez, Louis Faust, and Sidney K D'Mello. 2022. Toward robust stress prediction in the age of wearables: Modeling perceived stress in a longitudinal study with information workers. *IEEE Transactions on Affective Computing* 13, 4 (2022), 2201–2217.
[12] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.
[13] Adam Brown, Sarah D'Angelo, Ben Holtz, Ciera Jaspan, and Collin Green. 2023. Using logs data to identify when software engineers experience flow or focused work. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–12.
[14] Emily A Butler, Tiane L Lee, and James J Gross. 2009. Does expressing your emotions raise or lower your blood pressure? The answer depends on cultural context. *Journal of Cross-Cultural Psychology* 40, 3 (2009), 510–517.
[15] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.
[16] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2* 1, 4 (2015), 1–4.
[17] Ji Young Cho and Eun-Hee Lee. 2014. Reducing confusion about grounded theory and qualitative content analysis: Similarities and differences. *Qualitative*

report 19, 32 (2014).

[18] Seong-Sik Cho, Hyunjoo Kim, JinWoo Lee, Sinye Lim, and Woo Chul Jeong. 2019. Combined exposure of emotional labor and job insecurity on depressive symptoms among female call-center workers: A cross-sectional study. *Medicine* 98, 12 (2019).

[19] Eun Kyoung Choe, Saeed Abdullah, Mashfiqui Rabbi, Edison Thomaz, Daniel A Epstein, Felicia Cordeiro, Matthew Kay, Gregory D Abowd, Tanzeem Choudhury, James Fogarty, et al. 2017. Semi-automated tracking: a balanced approach for self-monitoring applications. *IEEE Pervasive Computing* 16, 1 (2017), 74–84.

[20] Vedant Das Swain, Victor Chen, Shrija Mishra, Stephen M Mattingly, Gregory D Abowd, and Munmun De Choudhury. 2022. Semantic gap in predicting mental wellbeing through passive sensing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–16.

[21] Vedant Das Swain, Lan Gao, Abhirup Mondal, Gregory D Abowd, and Munmun De Choudhury. 2024. Sensible and Sensitive AI for Worker Wellbeing: Factors that Inform Adoption and Resistance for Information Workers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–30.

[22] Vedant Das Swain, Lan Gao, William A Wood, Srikruthi C Matli, Gregory D Abowd, and Munmun De Choudhury. 2023. Algorithmic power or punishment: Information worker perspectives on passive sensing enabled ai phenotyping of performance and wellbeing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–17.

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

[24] Virginia Doellgast and Sean O'Brady. 2020. Making call center jobs better: The relationship between management practices and worker stress. (2020).

[25] Bryan Dosono and Bryan Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–13.

[26] Gerard Dunleavy, Ram Bajpai, André Comiran Tonon, Kei Long Cheung, Thuan-Quoc Thach, Yuri Rykov, Chee-Kiong Soh, Hein de Vries, Josip Car, and Georgios Christopoulos. 2020. Prevalence of psychological distress and its association with perceived indoor environmental quality and workplace factors in under and aboveground workplaces. *Building and environment* 175 (2020), 106799.

[27] Sidney K D'Mello and Brandon M Booth. 2023. Affect detection from wearables in the "real" wild: Fact, fantasy, or somewhere in between? *IEEE Intelligent Systems* 38, 1 (2023), 76–84.

[28] Don Samitha Elvitigala, Philipp M Scholl, Hussel Suriyaarachchi, Vipula Dissanayake, and Suranga Nanayakkara. 2021. StressShoe: a DIY toolkit for just-in-time personalised stress interventions for office workers performing sedentary tasks. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction*. Association for Computing Machinery, New York, NY, USA, 1–14.

[29] Tiantian Feng and Shrikanth S Narayanan. 2020. Modeling behavioral consistency in large-scale wearable recordings of human bio-behavioral signals. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1011–1015.

[30] Allison S Gabriel and James M Diefendorff. 2015. Emotional labor dynamics: A momentary approach. *Academy of management Journal* 58, 6 (2015), 1804–1825.

[31] Erving Goffman. 2016. The presentation of self in everyday life. In *Social Theory Re-Wired*. Routledge, 482–493.

[32] Alicia A Grandey. 2000. Emotional regulation in the workplace: A new way to conceptualize emotional labor. *Journal of occupational health psychology* 5, 1 (2000), 95.

[33] James J Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of general psychology* 2, 3 (1998), 271–299.

[34] Ted Grover, Kael Rowan, Jina Suh, Daniel McDuff, and Mary Czerwinski. 2020. Design and evaluation of intelligent agent prototypes for assistance with focus and productivity at work. In *Proceedings of the 25th international conference on intelligent user interfaces*. Association for Computing Machinery, New York, NY, USA, 390–400.

[35] Omar Hahad, Marin Kuntic, Sadeer Al-Kindi, Ivana Kuntic, Donya Gilan, Katja Petrowski, Andreas Daiber, and Thomas Münzel. 2024. Noise and mental health: evidence, mechanisms, and consequences. *Journal of Exposure Science & Environmental Epidemiology* (2024), 1–8.

[36] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications* 13, 4 (1998), 18–28.

[37] Javier Hernandez, Rob R Morris, and Rosalind W Picard. 2011. Call center stress recognition with person-specific models. In *Affective Computing and Intelligent Interaction: 4th International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part I 4*. Springer, 125–134.

[38] Javier Hernandez, Pablo Paredes, Asta Roseway, and Mary Czerwinski. 2014. Under pressure: sensing stress of computer users. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 51–60.

[39] Arlie Russell Hochschild. 1979. Emotion work, feeling rules, and social structure. *American journal of sociology* 85, 3 (1979), 551–575.

[40] Arlie Russell Hochschild. 2019. *The managed heart: Commercialization of human feeling*. University of California press.

[41] Karen Hovsepian, Mustafa Al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. Association for Computing Machinery, New York, NY, USA, 493–504.

[42] Esther Howe, Jina Suh, Mehrab Bin Morshed, Daniel McDuff, Kael Rowan, Javier Hernandez, Marah Ihab Abdin, Gonzalo Ramos, Tracy Tran, and Mary P Czerwinski. 2022. Design of Digital Workplace Stress-Reduction Intervention Systems: Effects of Intervention Type and Timing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 327, 16 pages. doi:10.1145/3491102.3502027

[43] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. 2020. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678* (2020).

[44] Houtan Jebelli, Byungjoo Choi, and SangHyun Lee. 2019. Application of wearable biosensors to construction sites. I: Assessing workers' stress. *Journal of Construction Engineering and Management* 145, 12 (2019), 04019079.

[45] Shiyi Jiang, Farshad Firouzi, Krishnendu Chakrabarty, and Eric B Elbogen. 2021. A resilient and hierarchical IoT-based solution for stress monitoring in everyday settings. *IEEE Internet of Things Journal* 9, 12 (2021), 10224–10243.

[46] Gyuwon Jung and Uichin Lee. 2025. CounterStress: Enhancing Stress Coping Planning through Counterfactual Explanations in Personal Informatics. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.

[47] Gyuwon Jung, Sangjun Park, and Uichin Lee. 2024. Deepstress: supporting stressful context sensemaking in personal informatics systems using a quasi-experimental approach. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[48] Soowon Kang, Woohyeok Choi, Cheul Young Park, Narae Cha, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Heepyung Kim, Yong Jeong, and Uichin Lee. 2023. K-emophone: A mobile and wearable dataset with in-situ emotion, stress, and attention labels. *Scientific data* 10, 1 (2023), 351.

[49] Anna Kawakami, Shreya Chowdhary, Shamsi T Iqbal, Q Vera Liao, Alexandra Olteanu, Jina Suh, and Koustuv Saha. 2023. Sensing wellbeing in the workplace, why and for whom? envisioning impacts with organizational stakeholders. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–33.

[50] Zachary D King, Judith Moskowitz, Begum Egilmez, Shibo Zhang, Lida Zhang, Michael Bass, John Rogers, Roozbeh Ghaffari, Laurie Wakschlag, and Nabil Alshurafa. 2019. Micro-stress EMA: A passive sensing framework for detecting in-the-wild stress in pregnant mothers. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 3, 3 (2019), 1–22.

[51] Uichin Lee, Joonwon Lee, Minsam Ko, Changhun Lee, Yuhwan Kim, Subin Yang, Koji Yatani, Gahgene Gweon, Kyong-Mee Chung, and Junehwa Song. 2014. Hooked on smartphones: an exploratory study on smartphone overuse among college students. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2327–2336.

[52] Boning Li and Akane Sano. 2020. Extraction and interpretation of deep autoencoder-based temporal features from wearables for forecasting personalized mood, health, and stress. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–26.

[53] Kun Liang, Anfu Zhou, Zhan Zhang, Hao Zhou, Huadong Ma, and Chenshu Wu. 2023. mmStress: Distilling human stress from daily activities via contact-less millimeter-wave sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–36.

[54] Wei-Yin Loh. 2011. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery* 1, 1 (2011), 14–23.

[55] Tom Lovett, Eamonn O'Neill, James Irwin, and David Pollington. 2010. The calendar as a sensor: analysis and improvement using data fusion with social networks and location. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. Association for Computing Machinery, New York, NY, USA, 3–12.

[56] Xi Lu, Edison Thomaz, and Daniel A Epstein. 2022. Understanding People's Perceptions of Approaches to Semi-Automated Dietary Monitoring. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–27.

[57] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017), 4768–4777.

[58] Robert C MacCallum, Shaobo Zhang, Kristopher J Preacher, and Derek D Rucker. 2002. On the practice of dichotomization of quantitative variables. *Psychological methods* 7, 1 (2002), 19.

[59] Gloria Mark, Shamsi T Iqbal, Mary Czerwinski, Paul Johns, Akane Sano, and Yuliya Lutchyn. 2016. Email duration, batching and self-interruption: Patterns of email use on productivity and stress. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 1717–1728.

[60] Akhil Mathur, Marc Van den Broeck, Geert Vanderhulst, Afra Mashhadi, and Fahim Kawsar. 2015. Tiny habits in the giant enterprise: understanding the dynamics of a quantified workplace. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 577–588.

[61] Stephen M Mattingly, Julie M Gregg, Pino Audia, Ayse Elvan Bayraktaroglu, Andrew T Campbell, Nitesh V Chawla, Vedant Das Swain, Munmun De Choudhury, Sidney K D'Mello, Anind K Dey, et al. 2019. The tesserae project: Large-scale, longitudinal, in situ, multimodal sensing of information workers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–8.

[62] Lakmal Meegahapola, William Droz, Peter Kun, Amalia De Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, et al. 2023. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–32.

[63] Jared Minkel, Marisa Moreta, Julianne Muto, Oo Htaik, Christopher Jones, Mathias Basner, and David Dinges. 2014. Sleep deprivation potentiates HPA axis stress reactivity in healthy adults. *Health Psychology* 33, 11 (2014), 1430.

[64] Varun Mishra, Gunnar Pope, Sarah Lord, Stephanie Lewia, Byron Lowens, Kelly Caine, Sougata Sen, Ryan Halter, and David Kotz. 2020. Continuous detection of physiological stress with commodity hardware. *ACM transactions on computing for healthcare* 1, 2 (2020), 1–30.

[65] Leanne G. Morrison, Charlie Hargood, Veljko Pejovic, Adam W. A. Geraghty, Scott Lloyd, Natalie Goodman, Danius T. Michaelides, Anna Weston, Mirco Musolesi, Mark J. Weal, and Lucy Yardley. 2017. The Effect of Timing and Frequency of Push Notifications on Usage of a Smartphone-Based Stress Management Intervention: An Exploratory Trial. *PLOS ONE* 12, 1 (01 2017), 1–15.

[66] Mehrab Bin Morshed, Javier Hernandez, Daniel McDuff, Jina Suh, Esther Howe, Kael Rowan, Marah Abdin, Gonzalo Ramos, Tracy Tran, and Mary Czerwinski. 2022. Advancing the understanding and measurement of workplace stress in remote information workers from passive sensors and behavioral data. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, IEEE Computer Society, Los Alamitos, CA, USA, 1–8.

[67] Amir Muaremi, Bert Arnrich, and Gerhard Tröster. 2013. Towards measuring stress with smartphones and wearable devices during workday and sleep. *BioNanoScience* 3 (2013), 172–183.

[68] Elaine M Murtagh, Jacqueline L Mair, Elroy Aguiar, Catrine Tudor-Locke, and Marie H Murphy. 2021. Outdoor walking speeds of apparently healthy adults: A systematic review and meta-analysis. *Sports Medicine* 51 (2021), 125–141.

[69] Sameer Neupane, Mithun Saha, Nasir Ali, Timothy Hnat, Shahin Alan Samiei, Anandatirtha Nandugudi, David M Almeida, and Santosh Kumar. 2024. Momentary Stressor Logging and Reflective Visualizations: Implications for Stress Management with Wearables. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–19.

[70] Kerstin Norman, Tohr Nilsson, Mats Hagberg, Ewa Wigaeus Tornqvist, and Allan Toomingas. 2004. Working conditions and health among female and male employees at a call center in Sweden. *American journal of industrial medicine* 46, 1 (2004), 55–62.

[71] PB Pankajavalli, GS Karthick, and R Sakthivel. 2021. An efficient machine learning framework for stress prediction via sensor integrated keyboard data. *IEEE Access* 9 (2021), 95023–95035.

[72] Eunji Park, Yugyeong Jung, Inyeop Kim, and Uichin Lee. 2023. Charlie and the Semi-Automated Factory: Data-Driven Operator Behavior and Performance Modeling for Human-Machine Collaborative Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–16.

[73] Eunji Park, Duri Lee, Yunjo Han, James Diefendorff, and Uichin Lee. 2024. Hide-and-seek: Detecting Workers' Emotional Workload in Emotional Labor Contexts Using Multimodal Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–28.

[74] Aditya Ponnada, Jixin Li, Shirlene Wang, Wei-Lin Wang, Bridgette Do, Genevieve F Dunton, and Stephen S Intille. 2022. Contextual biases in microinteraction ecological momentary assessment (μEMA) non-response. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–24.

[75] Sergei Popov, Stanislav Morozov, and Artem Babenko. 2019. Neural oblivious decision ensembles for deep learning on tabular data. *arXiv preprint arXiv:1909.06312* (2019).

[76] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems* 31 (2018).

[77] Laavanya Rachakonda, Arham Kothari, Saraju P Mohanty, Elias Kougianos, and Madhavi Ganapathiraju. 2019. Stress-Log: An IoT-based smart system to monitor stress-eating. In *2019 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 1–6.

[78] Nafiul Rashid, Trier Mortlock, and Mohammad Abdullah Al Faruque. 2023. Stress detection using context-aware sensor fusion from wearable devices. *IEEE Internet of Things Journal* 10, 16 (2023), 14114–14127.

[79] Sasank Reddy, Deborah Estrin, and Mani Srivastava. 2010. Recruitment framework for participatory sensing data collections. In *International Conference on Pervasive Computing*. Springer, 138–155.

[80] Verónica Rivera-Pelayo, Angela Fessl, Lars Müller, and Viktoria Pammer. 2017. Introducing mood self-tracking at work: Empirical insights from call centers. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 1 (2017), 1–28.

[81] Kat Roemmich, Florian Schaub, and Nazanin Andalibi. 2023. Emotion AI at work: Implications for workplace surveillance, emotional labor, and emotional privacy. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–20.

[82] Georgina Russell and Stafford Lightman. 2019. The human stress response. *Nature reviews endocrinology* 15, 9 (2019), 525–534.

[83] Wendy Sanchez, Alicia Martinez, Yasmin Hernandez, Hugo Estrada, and Miguel Gonzalez-Mendoza. 2023. A predictive model for stress recognition in desk jobs. *Journal of Ambient Intelligence and Humanized Computing* 14, 1 (2023), 17–29.

[84] Akane Sano and Rosalind W Picard. 2013. Stress recognition using wearable sensors and mobile phones. In *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, IEEE Computer Society, Los Alamitos, CA, USA, 671–676.

[85] Hillol Sarker, Matthew Tyburski, Md Mahbubur Rahman, Karen Hovsepian, Moushumi Sharmin, David H Epstein, Kenzie L Preston, C Debra Furr-Holden, Adam Milam, Inbal Nahum-Shani, et al. 2016. Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 4489–4501.

[86] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*. Association for Computing Machinery, New York, NY, USA, 400–408.

[87] Elena Smets, Emmanuel Rios Velazquez, Giuseppina Schiavone, Imen Chakroun, Ellie D'Hondt, Walter De Raedt, Jan Cornelis, Olivier Janssens, Sofie Van Hoecke, Stephan Claes, et al. 2018. Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *NPJ digital medicine* 1, 1 (2018), 67.

[88] Mauricio Soto, Chris Satterfield, Thomas Fritz, Gail C Murphy, David C Shepherd, and Nicholas Kraft. 2021. Observing and predicting knowledge worker stress, focus and awakeness in the wild. *International Journal of Human-Computer Studies* 146 (2021), 102560.

[89] Vedant Das Swain, Qiuyue Zhong, Jash Rajesh Parekh, Yechan Jeon, Roy Zimmerman, Mary Czerwinski, Jina Suh, Varun Mishra, Koustuv Saha, Javier Hernandez, et al. 2024. AI on My Shoulder: Supporting Emotional Labor in Front-Office Roles with an LLM-based Empathetic Coworker.

[90] Kobiljon Toshnazarov, Uichin Lee, Byung Hyung Kim, Varun Mishra, Lismer Andres Caceres Najarro, and Youngtae Noh. 2024. SOSW: Stress Sensing with Off-the-shelf Smartwatches in the Wild. *IEEE Internet of Things Journal* 11, 12 (2024), 21527–21545.

[91] Roy Van Den Heuvel and Carine Lallemand. 2023. Personal Informatics at the Office: User-Driven, Situated Sensor Kits in the Workplace. In *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work*. Association for Computing Machinery, New York, NY, USA, 1–13.

[92] Tommi Vehviläinen, Harri Lindholm, Hannu Rintamäki, Rauno Pääkkönen, Ari Hirvonen, Olli Niemi, and Juha Vinha. 2016. High indoor CO2 concentrations in an office environment increases the transcutaneous CO2 level and sleepiness during cognitive work. *Journal of occupational and environmental hygiene* 13, 1 (2016), 19–29.

[93] Jacqueline C Vischer. 2007. The effects of the physical environment on job performance: towards a theoretical model of workspace stress. *Stress and health: Journal of the International Society for the Investigation of Stress* 23, 3 (2007), 175–184.

[94] Steffanie L Wilk and Lisa M Moynihan. 2005. Display rule" regulators": the relationship between supervisors and worker emotional exhaustion. *Journal of applied psychology* 90, 5 (2005), 917.

[95] Peder Wolkoff. 2018. Indoor air humidity, air quality, and health–An overview. *International journal of hygiene and environmental health* 221, 3 (2018), 376–390.

[96] Jia Xu, Teng Xiao, Pin Lv, Zhe Chen, Chao Cai, Yang Zhang, and Zehui Xiong. 2024. Tracing human stress from physiological signals using UWB radar. *IEEE Internet of Things Journal* 11, 20 (2024), 32773–32790.

[97] Joanna C Yau, Benjamin Girault, Tiantian Feng, Karel Mundnich, Amrutha Nadarajan, Brandon M Booth, Emilio Ferrara, Kristina Lerman, Eric Hsieh,

and Shrikanth Narayanan. 2022. TILES-2019: A longitudinal physiologic and behavioral data set of medical residents in an intensive care unit. *Scientific Data* 9, 1 (2022), 536.

[98] Dieter Zapf, Amela Isic, Myriam Bechtoldt, and Patricia Blau. 2003. What is typical for call centre jobs? Job characteristics, and service interactions in different call centres. *European journal of work and organizational psychology* 12, 4 (2003), 311–340.

[99] Dieter Zapf, Christoph Vogt, Claudia Seifert, Heidrun Mertini, and Amela Isic. 1999. Emotion work as a source of stress: The concept and development of an instrument. *European Journal of work and organizational psychology* 8, 3 (1999), 371–400.

[100] Jingwen Zhang, Dingwen Li, Ruixuan Dai, Heidy Cos, Gregory A Williams, Lacey Raper, Chet W Hammill, and Chenyang Lu. 2022. Predicting post-operative complications with wearables: a case study with patients undergoing pancreatic surgery. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–27.

[101] Manuela Züger, Christopher Corley, André N Meyer, Boyang Li, Thomas Fritz, David Shepherd, Vinay Augustine, Patrick Francis, Nicholas Kraft, and Will Snipes. 2017. Reducing interruptions at work: A large-scale field study of flow-light. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 61–72.

[102] Manuela Züger, Sebastian C Müller, André N Meyer, and Thomas Fritz. 2018. Sensing interruptibility in the office: A field study on the use of biometric and computer interaction sensors. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–14.

A Appendix

A.1 Challenges During Data Collection

While call center agents' tasks are repetitive, the variability in customer needs and issues makes them highly dynamic. We collected data in actual settings rather than controlled environments to reflect these real-world conditions. Table 6 summarizes the steps, challenges, considerations, and actions taken during this process. This summary highlights the rationale behind optimizing our data collection decisions for real-world conditions in call centers and provides practical insights for future researchers.

A.2 Modeling Process for Quantitative Analysis

A.2.1 Data Cleaning. As a first step, we excluded three invalid participants based on post hoc investigations conducted after data collection. The two participants were identified as potentially providing insincere self-reported labels during the initial screening phase. Specifically, one participant labeled "little stress" (value 2) in more than 97.3% of 997 responses, and another labeled "no stress" (value 1) in more than 97.1% of 1,097 responses. While such distributions may reflect individual differences in real-world self-reports, they also raise concerns about response sincerity, particularly given the emotionally demanding work environment where call agents often interact with irate customers. After consultation with their managers, we determined that these participants' highly skewed response patterns were likely atypical and excluded them from further analysis. We also excluded one participant due to self-reported concerns regarding the temporal consistency of her own data, indicating perceived fluctuations in emotional state and uncertainty about the reliability of her responses over time. Second, we removed all data entries that occurred outside each participant's official working hours. As a result, we used the remaining data from 15 call agents (14 female, one male), spanning 330 days, 13,925 call logs, and 13,190 self-reported labels. We then conducted data quality checks on each sensor modality.

Data from the call log server. Two call center managers manually reviewed the call log data before sharing it. For initial data cleaning, the managers filtered the data approved by the organization (e.g., call start time, call duration) and excluded any personally identifiable information from the call agents' written summaries of the interactions, which consist of a pair of questions and answers. Then, researchers checked the dataset and removed duplicate records, irrelevant entries (e.g., records with a duration of 0), and call logs after obtaining participants' consent to use them. After these steps, a call log dataset (13,735 entries) was ready for analysis.

Data from Fitbit. To ensure device operation, we first verified the continuity of step count data, which is automatically recorded at one-minute intervals when the device is functioning (Fitbit Developer Guide). The results confirmed consistent operation throughout the study period. We then cleaned the HR and step data. Missing HR values, often due to non-wear or device misplacement [100], were retained as nulls given their low proportion (<3%). For step data, we removed entries where HR was also missing (3.78%) and identified outliers using an isolation forest algorithm [4]. Less than 0.2% of step entries were flagged and removed. We additionally verified that no step counts exceeded typical walking speed thresholds (115 steps/min) [68].

Data from the Tablet ACC and environmental sensors. We excluded 19 days when the ACC predominantly showed zeros or multiple entries with identical timestamps due to temporary data collection anomalies. Subsequently, on the ACC data, we implemented winsorization for each individual based on the three median absolute deviations (MAD) to address outliers. Specifically, we set the lower and upper bounds at Median -3*MAD and Median +3*MAD, respectively. Any values beyond these bounds were adjusted to the nearest boundary value. For environmental data (i.e., CO_2, temperature, and humidity), we excluded entries that showed zeros (18 days). The environmental data were used without additional processes.

A.2.2 Feature Extraction. We implemented two distinct windowing strategies for feature extraction: task- and time-based windows (Table 2). Table 7 describes the meaning of each feature.

A.2.3 Label Preprocessing. We first validated the self-reported stress labels. Matching call-termination times in call logs to self-reported stress labels produced 11,622 pairs; any label recorded after the next call began was discarded. To identify potentially insincere responses, we performed additional reliability checks on labels using other after-call survey questionnaires [64]. An additional screen removed 1,769 responses with identical ratings across all four after-call items (arousal, valence, surface acting, deep acting), yielding 9,853 reliable labels. Then, the 5-point Likert scale responses were binarized per participant to mitigate the influence of the interpersonal difference on stress labels [37]. Responses below the personal average were classified as 'low stress,' and those above were classified as 'high stress' [50].

A.2.4 Dataset Integration. After data cleaning and label encoding, we integrated multimodal features (e.g., accelerometer, Fitbit, environmental sensor) with corresponding stress labels. We constructed seven feature sets—one task-based and six time-based (with varying window sizes)—based on the extracted features. We removed all first-day data for participants to allow the experiment time to adjust to the setup and response methodology [51]. After filtering for any missing values from the integrated dataset, a total of 7,442 valid entries were retained for analysis. The final data distribution across 15 participants covered 310 days (mean = 20.67, SD = 2.58), with 66.10% labeled as 'low stress' and 33.90% as 'high stress.'

A.2.5 Normalization and Categorical Feature Preprocessing. To improve generalizability and reduce participant-specific bias, numerical features were z-score normalized. Missing values were removed before normalization. Categorical features were either one-hot encoded or used directly, depending on the model's input requirements (e.g., CatBoost). All preprocessing steps were applied separately to the training and evaluation sets.

A.2.6 Model Training and Evaluation.

Feature selection. Feature selection (i.e., LASSO [71]) was applied only to ML models that are sensitive to redundant or correlated input features. We optimized the L1 regularization parameter α through a grid-search approach, selecting the value that maximized the area under the ROC curve (AUC-ROC). α was chosen from 50

Table 6: Detailed procedure of data collection

Stage	Activities	Challenges	Considerations	Action Taken
Organization On-boarding	Collaborating with the organization's representative (manager) to design data collection method	1) Restrictions on data types and sensors imposed by regulations 2) Limitations on integrating tools with call center operating systems 3) Concerns about potential impact of data collection on work performance	1) Identification of allowable data types and sensors within institutional guidelines 2) Security regulations for setting up a network to collect data 3) Designing self-report data collection methods with minimal disruption	1) Review of institutional policies to determine permissible data types and sensors 2) Designing and implementing a data collection procedure with separate devices 3) Selection of concise survey questions followed by a pilot study to evaluate its impact on performance
Participant Recruitment	Hosting informational sessions to explain the data collection process and gather feedback from potential participants	1) Privacy concerns related to mobile data collection 2) Concerns about long-term use of unfamiliar wearable devices 3) Concerns about workload transfer to non-participants	1) Reassessment of mobile data collection 2) Considering participant comfort to decide smartwatch 3) Discussing ways to alleviate non-participants' concerns with the manager	1) Limitation on mobile data collection and usage 2) Evaluation of smartwatch devices and adoption of Fitbit for its convenience 3) Dispelling the concerns by sharing the results of a pilot study
Consent Procedure	Obtaining consent from participating call agents and each customer at every call	1) Complexity of securing consent from phone consultation customers	1) Legal validity to collect user consent of customer 2) Methods to compromise customer convenience 3) Addressing expected actions on additional consent processes (i.e., decline to listen to user consent)	1) Conducted legal review on the necessity of consent 2) Development of a concise and legally compliant consent statement 3) Modification of the system to incorporate the consent to confirm their consent immediately after the call connection and categorizing their responses into consent, non-consent, and no selection
Data Collection	Keeping engagement motivated and monitoring	1) Lack of on-site support staff for data collection 2) Monitoring and error handling during data collectio 3) Maintaining participant motivation	1) Remote data collection and monitoring system 2) Establishment of error response protocols 3) Creation of an environment that fosters participant motivation	1) Hiring a remote monitoring staff for daily oversight 2) Implementation of a two-step error resolution process: self-adjustment by participants, visiting equipment replacement 3) Provision of team-wide encouragement (e.g., snacks for all workers)

Table 7: Feature definition and category for task-based window

Name	Definition	Category explanation
Customer interaction (CI) duration	Duration of call recorded on server	-
Non-customer interaction (nCI) duration	Duration between calls calculated based on each call start time (In case of the first call in a day, duration between time to record baseline survey and the start time of call)	-
Inquiry and Answer length	Number of characters in records documented by call agents on server (In case of space, counting as one character)	-
Mute for CI	Time difference between length of recorded audio file and call duration time on the server (In case of call agent using mute button, server stopped record)	0: not used, 1: used
Agreement type for CI	The type of customer consent obtained for providing call voice recording information before the call	0: disagreement, 1: agreement, 2: else
Complaint type for CI	Type of content manually evaluated by the call agents as the internal criteria	0: compliment, 1: complaint, 2: else
Eating	Consumption of food between the end of the previous call and the end of the current call	1: eat, 0: not eat
Drinking	Consumption of drink between the end of the previous call and the end of the current call	1: drink, 0: not drink
Weekday	Day of the week, the data was collected	1: Monday, 2: Tuesday, 3: Wednesday, 4: Thursday, 5: Friday, 6: Saturday, 7: Sunday
Category of hour	Time slot during the day when the data was collected	0: before 8:00 a.m, 1: 8:00 a.m.-10:00 a.m., 2: 10:00 a.m.-12:00 p.m., 3: 12:00 p.m.-2:00 p.m., 4: 2:00 p.m.-4:00 p.m., 5: 4:00 p.m.-6:00 p.m., 6: 6:00 p.m.-9:00 p.m.

Table 8: Change of threshold (<) by the cut

pnum	mean	50% (median)	40%
1	1.576 (1)	1	1
2	2.265 (2)	2	2
3	1.430 (1)	1	1
4	1.243 (1)	1	1
5	1.571 (1)	1	1
6	1.173 (1)	1	1
7	2.628 (2)	2	2
8	3.200 (2)	3	2
9	1.768 (1)	1	1
10	1.516 (1)	1	1
11	1.891 (1)	2	2
12	2.235 (2)	2	1
13	1.150 (1)	1	1
14	1.438 (1)	1	1
15	1.967 (1)	1	1

values, logarithmically spaced between 10^{-4} and 10^{1}, and the model was trained with a maximum of 1000 iterations. All processing was performed in Python using the scikit-learn library. Extra feature selection was not performed for gradient-boosted decision tree models, which inherently perform feature selection and regularization as part of their hyperparameter optimization during training. Similarly, DL models were not manually restricted, as this could limit the model's ability to learn meaningful representations from heterogeneous modalities.

Oversampling. Depending on the model and feature format, we used either SMOTE [15] or SMOTENC before model training, applied separately to each participant's data to preserve intra-individual structure. SMOTE and SMOTENC generate synthetic samples by interpolating between the k-nearest neighbors ($k = 5$ separately). All procedures were implemented using the imbalanced-learn package in Python.

Models for training and evaluation. We utilized six ML models: Decision Tree (DT) [54], Linear Discriminant Analysis (LDA) [9], Support Vector Machine (SVM) [36], Random Forest (RF) [12], eXtreme Gradient Boosting (XGBoost) [16] and CatBoost [76]; and three DL models: Tabnet [6], Tabtransformer [43], and NODE [75]. For the ML models, we used the scikit-learn library and XGBoost package in Python. For the XGBoost, we used the default values (i.e., learning rate of 0.1 and a max depth of 3). We set the number of base estimators for three ensemble models (i.e., RF and XGBoost) to 100 [102], and the minimum sample split for tree-based models (i.e., DT and RF) to 50. For the CatBoost, we used the default values (i.e., learning rate of 0.03, iteration = 1000, and depth = 6). The hyperparameters of all DL models were tuned via Hyperopt. A batch size of 512 was adopted for DL models based on experimental results using Tabnet.

Training and evaluation. Training data were shuffled before each epoch, and batch normalization was omitted to avoid mixing samples from different participants within a batch. The max evaluation for hyperparameter tuning was limited to 20 per model, and the configuration yielding the highest validation AUC-ROC was selected. Also, we employed majority voting as a baseline to validate model performance, as this was our initial attempt to build a model. To evaluate the trained model's performance, we used the leave-one-subject-out (LOSO) cross-validation method [23, 86] as a default evaluation method. Model performance was measured using AUC-ROC and weighted F1-score, both of which are robust to class imbalance. The Wilcoxon signed-rank test ($\alpha = 0.05$) was used to assess statistical significance between paired results. Each test dataset is tested 30 times with a fixed random seed for each turn, and we report the results by calculating the average performance.

A.3 Evaluation on the Robustness of Label Binarization Strategies

We used each participant's mean value as a threshold for binarization of label to mitigate interpersonal difference. While dichotomizing using mean value is the common methods for affect sensing [50], it can reduce information and introduce threshold artifacts [5, 58]. To assess robustness, we repeated the exact same training and evaluation pipeline under multiple participant-specific thresholds.

Threshold definitions. For each participant i, we computed a participant-specific threshold T_i from their 1–5 post-call stress ratings using (a) the mean, (b) the median (50th percentile), and (c) a quantile-based cut (40th percentile). We then defined the binary label as $y^{bin} = 1$ if $y > T_i$ (high stress), and 0 otherwise for (a)–(c).[3] Table 8 summarizes how the resulting thresholds vary across participants under each strategy.

Evaluation protocol and metrics. For each thresholding strategy, we re-train and evaluate the exact same pipeline using two feature extraction strategies (i.e., task-alinged, and 5-min time-based) and two ML models (i.e., RandomForest, and XGBoost). We report ROC-AUC, PR-AUC, weighted F1, and class-wise precision/recall averaged across LOSO folds (mean ± SD).

Results. While adjusting the thresholds led to trade-offs in specific metrics such as Recall, the overall discriminative power of the model (ROC-AUC) remained stable without significant fluctuations (Table 9). Numerical fluctuations were observed in Recall_1 and Weighted F1-score, the Wilcoxon signed-rank test indicated no statistically significant difference, suggesting that the model's performance remains consistent across different thresholds.

A.4 Statistical Analysis on Windowing Strategies

Using a task-aligned and 5-min time-based feature resulted in a similar performance. A statistical analysis (Table 10) revealed no statistically significant differences between the two.

A.5 Ablation Study to Compare Importance of Data Type

Table 11 includes the results of every case in the ablation study.

[3]Because ratings are discrete (1–5), quantile-based cut points can collapse to the same integer threshold for some participants.

Table 9: Model performance by label encoding strategies

Task-aligned_RF	ROC-AUC	PR-AUC	Weighted F1-score	precision_0	recall_0	precision_1	recall_1
mean	0.68506 (0.09027)	0.51373 (0.19675)	0.61707 (0.16809)	0.75510 (0.15870)	0.73214 (0.24079)	0.51269 (0.23043)	0.49256 (0.21868)
50 % (median)	0.70736* (0.07775)	0.49635 (0.18747)	0.69958 (0.07206)	0.78018* (0.09345)	0.82154* (0.10734)	0.50901 (0.19627)	0.42233* (0.16649)
40 %	0.69658 (0.08679)	0.51482 (0.20379)	0.68828 (0.07724)	0.75284 (0.13353)	0.79996 (0.1123)	0.51445 (0.2076)	0.44633 (0.16025)
Task-aligned_XGBoost	ROC-AUC	PR-AUC	Weighted F1-score	precision_0	recall_0	precision_1	recall_1
mean	0.67902 (0.08934)	0.51580 (0.06940)	0.60100 (0.14500)	0.74247 (0.25287)	0.73243 (0.18299)	0.53455 (0.24142)	0.46543 (0.16355)
50 % (median)	0.69940* (0.08101)	0.49821 (0.18694)	0.67321 (0.08340)	0.77333 (0.10406)	0.82153* (0.13927)	0.52671 (0.22099)	0.39343 (0.21546)
40 %	0.69212 (0.07880)	0.52372 (0.18901)	0.66867 (0.08169)	0.74173 (0.13691)	0.80539 (0.15108)	0.54075 (0.23904)	0.41157 (0.17996)
5-min time-based_RF	ROC-AUC	PR-AUC	Weighted F1-score	precision_0	recall_0	precision_1	recall_1
mean	0.69172 (0.08168)	0.51800 (0.19736)	0.63077 (0.12823)	0.75666* (0.15175)	0.72651* (0.20712)	0.51309 (0.22256)	0.51769* (0.21045)
50 % (median)	0.70510 (0.07973)	0.49706 (0.18333)	0.69488 (0.07191)	0.78302 (0.09596)	0.80394 (0.11455)	0.50147 (0.20500)	0.45434 (0.14781)
40 %	0.69679 (0.08281)	0.51726 (0.20209)	0.68557 (0.07696)	0.75516 (0.13431)	0.78708 (0.12846)	0.52014 (0.21459)	0.46938* (0.1468)
5-min time-based_XGBoost	ROC-AUC	PR-AUC	Weighted F1-score	precision_0	recall_0	precision_1	recall_1
mean	0.68637 (0.08805)	0.52488 (0.19205)	0.61003 (0.14233)	0.73843 (0.15584)	0.74725 (0.20522)	0.50392 (0.22642)	0.46384 (0.23572)
50 % (median)	0.70565(0.08255)	0.50950 (0.18028)	0.67185 (0.08185)	0.77168 (0.10731)	0.81818* (0.13053)	0.51678 (0.21512)	0.40720 (0.2051)
40 %	0.70191 (0.07856)	0.52923 (0.18582)	0.65937 (0.07646)	0.74492 (0.13760)	0.78930 (0.14612)	0.52117 (0.22562)	0.43444 (0.19208)

Note: Values are reported as Mean (Standard Deviation). * Indicates a statistically significant difference compared to the **'mean'** method at the $p < 0.05$ level. Statistical significance was determined using a two-sided **Wilcoxon signed-rank test** on paired samples.

Table 10: Comparison of RandomForest clasification performance between task and 5-min methods (Wilcoxon Signed-Rank Test)

Metric	**Task Mean**	**5-min Mean**	***p*-value**	**Result**
AUC	0.6851	0.6917	0.1876	n.s.
PR AUC Score	0.5137	0.5180	0.4543	n.s.
Weighted F1-score	0.6171	0.6308	0.8040	n.s.

Note: n.s. denotes not significant ($p > 0.05$).

Table 11: Ablation study (task-aligned window, RF)

Passive sensing	Self-reported	ROC-AUC	PR-AUC	Weighted F1-score	precision_0	recall_0	precision_1	recall_1
All	-	0.68896 (0.08557)	0.51926 (0.19560)	0.62486 (0.15904)	0.75041 (0.15545)	0.73650 (0.22674)	0.51386 (0.21792)	0.49957 (0.2233)
Task	-	0.69763 (0.07883)	0.53085 (0.53085)	0.66447 (0.09605)	0.75203 (0.15189)	0.76962 (0.09930)	0.51265 (0.21265)	0.50218 (0.13676)
Task + Env.	-	0.68080 (0.09011)	0.51617 (0.20126)	0.65165 (0.11531)	0.73844 (0.14912)	0.78009 (0.11408)	0.50519 (0.22181)	0.44122 (0.14398)
Task + Env. + Behavioral	-	0.69005 (0.08494)	0.52283 (0.19194)	0.63011 (0.15401)	0.75182 (0.15406)	0.73793 (0.21770)	0.51277 (0.21125)	0.50254 (0.21469)
Without task	-	0.66186 (0.07948)	0.48158 (0.19090)	0.60900 (0.15757)	0.72488 (0.15757)	0.68680 (0.23345)	0.47719 (0.24055)	0.46986 (0.15793)
All	A	0.68940 (0.08521)	0.52029 (0.19590)	0.62625 (0.15810)	0.74929 (0.15530)	0.73811 (0.22620)	0.51458 (0.21654)	0.49726 (0.22070)
Task	A	0.69594 (0.07855)	0.52902 (0.18622)	0.66392 (0.09804)	0.77183 (0.09762)	0.74295 (0.08671)	0.51239 (0.21129)	0.49756 (0.13471)
Task + Env.	A	0.67956 (0.09210)	0.51574 (0.20194)	0.65216 (0.11649)	0.73867 (0.14946)	0.78028 (0.11248)	0.50345 (0.22071)	0.44068 (0.14596)
Task + Env. + Behavioral	A	0.69005 (0.08514)	0.52167 (0.19237)	0.62823 (0.15660)	0.75260 (0.15392)	0.73666 (0.22152)	0.51246 (0.21149)	0.50122 (0.21648)
Without task	A	0.65639 (0.08189)	0.47255 (0.19021)	0.61030 (0.14678)	0.73072 (0.15166)	0.71532 (0.22701)	0.47521 (0.18802)	0.46953 (0.23440)
All	A + B	0.68945 (0.08573)	0.51890 (0.19477)	0.62224 (0.15451)	0.75181 (0.15466)	0.73203 (0.22723)	0.51024 (0.21494)	0.50122 (0.22092)
Task	A + B	0.69061 (0.07915)	0.52316 (0.18846)	0.64627 (0.11062)	0.74363 (0.15156)	0.76807 (0.12892)	0.51006 (0.21847)	0.47399 (0.12892)
Task + Env.	A + B	0.68004 (0.09129)	0.51440 (0.19922)	0.64621 (0.12081)	0.77298 (0.14923)	0.73373 (0.13507)	0.50842 (0.21788)	0.45129 (0.14904)
Task + Env. + Behavioral	A + B	0.68737 (0.08581)	0.51680 (0.19318)	0.62170 (0.15987)	0.75054 (0.15333)	0.72970 (0.23162)	0.50878 (0.21273)	0.50139 (0.21942)
Without task	A + B	0.65474 (0.08295)	0.47089 (0.19056)	0.59926 (0.15511)	0.72782 (0.15779)	0.70435 (0.24732)	0.46799 (0.19281)	0.47617 (0.24677)

Note: A in self-reported data means self-reported after-call data (i.e., food intake), and B does self-reported daily baseline (e.g., arousal before work).