scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A Multimodal Dataset for Assessing Emotion, Stress, and Emotional Workload in Interpersonal Work Scenario

Duri Lee[1], Eunji Park[2], Gyuna Kim[3], Yunjo Han[4] & Uichin Lee[1]✉

We present EmoWork, a multimodal, multi-label dataset designed to support emotion and stress detection in realistic interpersonal work settings. Interpersonal work—common in occupations such as customer service—often requires workers to regulate their emotional expressions in response to strong affective stimuli. These demands, shaped by organizational display rules, present a unique challenge for affective computing systems, particularly in scenarios where internal emotional states diverge from observable behaviors. Despite this, no public datasets exist that capture such dynamics of affect in naturalistic settings. To address this gap, we collected physiological, behavioral, and self-reported data from call center workers who engaged in role-play scenarios simulating customer service interactions with professional actors portraying dissatisfied customers. The dataset includes self-reported affective ratings, which are used as labels for classification, synchronized recordings from three wearable devices (i.e., Polar H10, Empatica E4, and Muse S), and features extracted from video and audio data. The EmoWork dataset advances affective computing by offering context-rich, multimodal data grounded in realistic interpersonal work scenarios.

Background & Summary

Affective computing focuses on creating systems that can recognize, interpret, and understand *human* affect (i.e., emotion and stress)[1]. These systems serve as an affect-aware foundation for *human* interaction-centered disciplines. Examples include *Human*-Computer Interaction (HCI)[2], *Human*-Robot Interaction (HRI)[3,4], and *Human*-AI Interaction (HAI)[5] across everyday contexts, such as home[6], workplace[7,8], and driving[9]. Among these contexts, the workplace has emerged as a particularly prominent application domain, as Industry 4.0 has transformed it into an environment rich in sensing technologies and driven by digital innovation[10,11]. Within work environments, researchers have extensively studied automatic monitoring of workers' mental status, such as workload[12], stress[13,14], and mood[15], using diverse sensors (e.g., computer logs[16], mobile[17], and wearable sensors[18]) to enhance workers' well-being and performance.

Prior studies relied on automatic affect-detection models, whose performance depends on securing relevant datasets. To acquire reliable datasets in workplace settings, it is important to reflect different contextual information such as the type of work (e.g., construction, service work, knowledge work, and manufacturing) because affective responses are highly context-dependent[19,20]. However, most publicly available datasets largely focused on capturing responses to standardized affective stimuli (e.g., watching videos[21–27] or Trier Social Stress Test (TSST)[28]) (Table 1). Although some studies have attempted to collect affective datasets in more naturalistic scenarios (e.g., knowledge worker scenario[29], interpersonal conversations[30,31]), these efforts still fall short in a diversity of scenarios from work types. Especially, data on interpersonal workplace interactions (e.g., conversations between customers and agents) remain scarce in existing datasets for affective computing. Yet, such scenarios are not marginal as interpersonal work is central to the modern labor force.

Despite this gap, interpersonal work—occupations that involve direct interaction with clients or coworkers—constitutes over half of all global employment[32] and includes large job families such as the 2.9 million U.S.

[1]KAIST, School of Computing, Daejeon, 34141, South Korea. [2]Chung-Ang University, Department of Computer Science and Engineering, Seoul, 06974, South Korea. [3]KAIST, Graduate School of Data Science, Daejeon, 34141, South Korea. [4]Samsung Research, Samsung Electronics, Seoul, 06765, South Korea. ✉e-mail: uclee@kaist.ac.kr

Dataset	Condition	Collected Data Type	Self-reported Labels (frequency)	Participants
DEAP[21] (2012)	Stimulus (watching videos)	Physiological data (EEG, EDA, BVP, RESP, TEMP, EMG, EOG)	SAM and liking (after each stimulus)	32
SWELL-KW[29] (2014)	Scenario (conducting tasks of typical knowledge work)	Physiological data (ECG, EDA)	NASA-Task Load Index, SAM, and stress (after task)	25
DECAF[22] (2015)	Stimulus (watching videos)	Physiological data (MEG, hEOG, ECG, tEMG), Behavioral data (facial video)	Arousal and valence (after each stimulus)	30
ASCERTAIN[23] (2016)	Stimulus (watching videos)	Physiological data (ECG, EDA, EEG), Behavioral data (ACC)	Arousal, valence, engagement, liking, and familiarity (after each stimulus)	58
DREAMER[24] (2018)	Stimulus (watching videos)	Physiological data (EEG, ECG)	SAM (after each stimulus)	23
WESAD[28] (2018)	Stimulus (watching videos, TSST)	Physiological data (ECG, EDA, EMG, RESP, TEMP, BVP), Behavioral data (ACC)	PANAS, STAI, SAM, SSSQ, and four affect questions (5 times per session)	15
K-EmoCon[30] (2020)	Scenario (naturalistic conversation)	Physiological data (EEG, ECG, BVP, EDA, TEMP), Behavioral data (ACC)	Arousal, valence, and affective categories (continuous annotation per 5 seconds)	32
AMIGOS[25] (2021)	Stimulus (watching videos)	Physiological data (EEG, ECG, EDA)	SAM, liking, familiarity, self-assessment level, and Ekman's basic emotions (before and after each stimulus)	40
Emognition[26] (2022)	Stimulus (watching films)	Physiological data (EEG, BVP, HR, EDA, skin Temperature), Behavioral data (ACC, and GYRO, and the upper-body videos)	Nine emotions, and SAM (after each stimulus)	43
FACED[27] (2023)	Stimulus (watching videos)	Physiological data (EEG)	Nine emotions, arousal, valence, familiarity, and liking (after each stimulus)	123
UAV[31] (2024)	Scenario (cue-based conversation scenario)	Physiological data (EEG), Behavioral data (audio and video recording)	Arousal and valence (after each conversation)	42
EmoWork	Scenario (dyadic conversation in call center)	Physiological data (EEG, ECG, BVP, EDA, TEMP), Behavioral data (ACC, features from audio, features from video)	Arousal, valence, stress, and emotional workload (continuous annotation per 100 ms)	31

Table 1. Open datasets collected in lab environments for affective computing. This table summarizes publicly available datasets for studying affective computing, categorized by experimental conditions (stimulus-based or scenario-based). It includes the types of physiological and behavioral data collected, self-reported labels, and the number of participants for each dataset.

customer-service representatives (CSR), underscoring its prevalence in today's workforce[33]. These CSR roles, especially in the service sector, are emotionally demanding, as they frequently expose workers to strong affective stimuli while simultaneously requiring emotional suppression in accordance with organizational display rules[34]. Such dual demands, intense emotional elicitation, and strict regulation create a unique challenge for affect detection because workers' internal states often diverge from observable expressions. Therefore, realistic interpersonal work scenarios are essential not only for understanding these affective dynamics but also for training and evaluating affect-detection models that can operate under socially regulated conditions. One prominent CSR example is call center work[35,36], where agents regularly engage in emotionally charged conversations with dissatisfied customers. However, few publicly available datasets exist that authentically reflect interpersonal emotional workload. To fill this research gap, we built EmoWork, a context-specific multimodal dataset captured in realistic interpersonal work scenarios.

EmoWork is a multimodal, multi-label dataset from call agents engaged in real-time voice interactions with dissatisfied customers. It captures four synchronized affective states (i.e., stress, suppression, arousal, and valence), as well as physiological and behavioral responses. For realistic data collection in an interpersonal work setting, we employed scripted role-play scenarios simulating customer service interactions as a method frequently used in psychology literature on emotional labor[35]. We designed a scenario of a credit-card dispute, an issue known to provoke sensitive and challenging reactions from customers. The role-play scenario was designed to explicitly incorporate display rule compliance, a core aspect of emotional labor. To maximize realism, we hired trained actors to portray difficult customers. During the simulated calls, we measured physiological and behavioral responses using three wearable devices (i.e., Polar H10, Muse S, Empatica E4), a headset, and a camera. Following the calls, participants retrospectively reported their momentary levels of arousal, valence, stress, and emotional workload (i.e., suppression). Our dataset, collected in a realistic interpersonal work scenario with a high emotional workload, lays the groundwork for future research on affect modeling in such scenarios. In the following section, we describe the experimental protocol, sensor suite, and technical validation in detail.

Methods

Ethics Statement. We obtained Institutional Review Board (IRB) approval for open data collection and release from our institution (KAIST IRB) and conducted the study according to the approved protocol (KH2022-108). Recruitment was conducted through the managers of each local call center, who were contacted through institutional training coordinators. Participation was entirely voluntary, and those who expressed willingness to join were enrolled sequentially on a first-come, first-served basis. Before participation, all participants received a written explanation of the study, including the types of data to be collected (audio, video, physiological data, and self-reported responses), the purpose of the research, and how the collected data would be used and shared for academic purposes. Participants provided written informed consent acknowledging their understanding and voluntary agreement to data collection, anonymized data sharing, and the public release of derived datasets under

IRB-approved conditions. Also, female participants were asked whether they were pregnant, and pregnant women were excluded from the study. During data collection, participants could withdraw from the study at any time. After data collection, breaks were provided that incorporated meditation materials to relieve any negative effects on participants' mental well-being. To protect participants' information, all data were pseudo-anonymized during the data collection process. Call agents were assigned participant numbers as identifiers, and the actors used only fictitious personal details (e.g., names, phone numbers, and jobs) in the simulated call scenarios. All raw data were securely stored on a restricted-access shared drive managed by authorized researchers, and any information that could potentially identify individuals was strictly excluded from publication and public data release.

Data Collection. *Participants.* Our data collection process involved two distinct roles: customer and call agent. The customers were played by four professional actors from the local theater society (female: 2, male: 2). The call agents were recruited from real call agents (female: 24, male: 7) working in local call centers operated by six different companies. The gender distribution (77.4% female) aligns with prior research showing that call center work in South Korea is heavily female-dominated[37]. Their educational backgrounds included 19 with a high school diploma and 12 with a college degree or higher (mean age $= 36.19 \pm 11.38$). Because official statistics on call center demographics (e.g., education level and age) are limited and often inconsistent across government and industry reports, these values should be regarded as indicative rather than definitive. A total of 34 participants were recruited, but three dropped out before the data collection procedure (one was screened due to pregnancy; another due to COVID-19 infection; and the third withdrew voluntarily).

Data Collection Setup. The data collection process consisted of two phases: 1) sensor data collection during voice calls and 2) self-reported retrospective data collection after voice calls[35]. The second phase was carried out only with call agents, because this dataset was designed to collect data to evaluate their emotions and stress. Data collection was conducted in three rooms throughout the study, with one researcher assigned to each room, for a total of three researchers to assist with data collection.

Sensor data collection. Two separate rooms were assigned for the customer and the call agent. Each room was divided into a preparation and monitoring space and a data collection space using a partition. In the preparation and monitoring space, participants completed all pre-session procedures (e.g., providing explanations of the roles and procedures, and obtaining informed consent). During data collection, a researcher remained in the preparation and monitoring space and monitored the incoming data in real-time, checking for equipment malfunctions. The data collection space was set up differently depending on the participants' roles. Each participant involved in the role of call agent sat in front of a computer screen, wore a headset with a microphone, and answered customer calls using Skype (Fig. 1). The headset was physically connected to the desktop, and audio data was captured by the headset's microphone and stored in a designated folder on the desktop computer. This audio recording includes only the voice of call agents. A Logitech camera was installed on the desktop, connected via a USB cable. The recorded video was saved to a designated folder on the desktop. The call agent participants also wore a headband (i.e., Muse S), a chest band (i.e., Polar H10), and a wristband (i.e., Empatica E4) to collect multimodal physiological and behavioral data. Muse S collected electroencephalogram (EEG) data, which was recorded on a local computer via Bluetooth. The system was coded using the muse-lsl (https://pypi.org/project/muselsl/) Python package. Polar H10 collected electrocardiogram (ECG) data that was transferred via Bluetooth to a designated mobile data collection device (Galaxy Note 9). It was operated via Polar Beat (https://support.polar.com/en/beat). Empatica E4 worn on the non-dominant hand collected blood volume pulse (BVP), skin temperature (temp), inter-beat interval (IBI), electrodermal activity (EDA), and accelerometer (ACC), which were transmitted via Bluetooth to a mobile data collection device (Galaxy Note 9). Also, clear expected behaviors for participants—be nice to the customer and strive to satisfy the customer—were explicitly mentioned and posted on the wall in front of the call agent.

Each participant involved in the role of customers used a laptop placed inside a soundproof box. They used the same type of headset as the call agents, and their voices during phone calls were recorded in a designated folder on the laptop. For dialogue, customers initiate a call to a call agent using Skype, and the dyadic conversation files from Skype were saved as separate files. In addition, clear guidelines for customers in accordance with the required type of behavior were posted on the wall.

Self-reported data collection. We prepared a room for the call agent participants in the second phase, which was divided into a preparation space and a self-report data collection space. In the preparation space, the participants completed the post-data collection surveys and were instructed to store their personal belongings that could distract them to ensure participants' immersion during the retrospective self-reporting process. In the self-reported data collection space, participants sat in front of a laptop to record the label data measured on a Likert scale (Fig. 2). We used E-Prime 3.0, which has been widely used in psychology[38,39] with a customized interface to record participants' self-reported affective responses during each stimulus event (i.e., while listening to their own calls recorded in the previous session). Participants continuously reported their affective state by moving a mouse cursor on an interactive response screen (i.e., a slider with an anchor or a grid image with an anchor trackings the movements of the user), and the program recorded the cursor position every 100 ms[35]. This allowed us to capture continuous, time-varying self-report trajectories rather than discrete, one-time ratings. Both left- and right-handed participants used their right hand to operate the mouse.

Procedure. The overall procedure is illustrated in Fig. 3, and its details are explained in the following sections.

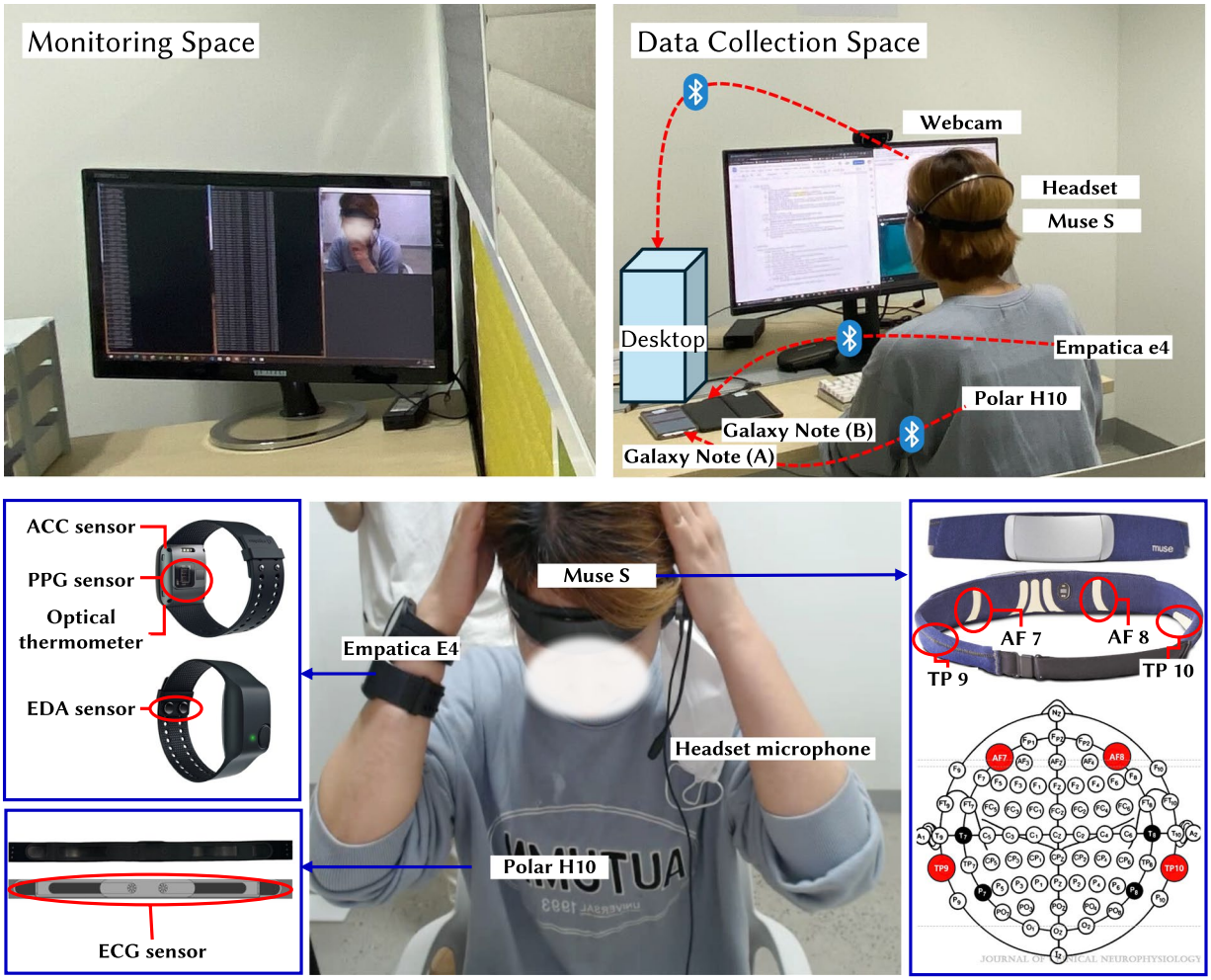


Fig. 1 Data collection setup. The room to collect data from call agents consisted of two spaces: a monitoring and data collection space. In the data collection space, the webcam and headset were wired to a desktop. Muse S sent signals to a desktop, Polar H10, and Empatica E4 sent signals to each assigned mobile device via Bluetooth. The call agents wore a headband (i.e., Muse S), headset, wristbands (i.e., Empatica E4 and Galaxy Watch 5), and a chestband (i.e., Polar H10).

Onboarding. We instructed call agent participants to use a pre-assigned subject number in every process for pseudonymization before arriving at the data collection site. Before data collection, they completed a pre-survey on basic demographic information (i.e., sex, age, work experience, education, and dominant hand) and baseline mental well-being such as the Patient Health Questionnaire (PHQ)-9[40], Perceived Stress Scale (PSS)[41], Positive Psychological Capital (PPC)[42], Big Five Inventory (BFI)[43], Rosenberg Self-Esteem Scale (RSE)[44], and Satisfaction with Life Scale (SWLS)[45]. All surveys were administered in Korean using validated Korean translations of the original instruments (Table 2). For the BFI, we used the 15-item Korean short version[46], which was developed and validated as a psychometrically reliable adaptation of the original 44-item scale. Participants were asked to avoid drinking alcohol, smoking, engaging in vigorous exercise, or consuming caffeine on the day of data collection, as these behaviors could affect their physiological responses. Upon arrival at the site, a researcher checked whether the pre-survey was completed and verified the subject number. The participants then read the study instructions and completed the user consent form.

A researcher in the designated space explained the role and provided supplementary documents. The documents included predefined opening and closing scripts for call agents, containing fictional information (i.e., name and call center). While the opening and closing scripts were predetermined, participants responded to the customer's requirements as they normally would at work, without a set script. Examples were provided to guide responses in predictable conflict situations. As in a real work environment, they did not have any prior information about the customer or the problem. Each participant practiced their role until they were familiar enough to begin data collection, and were allowed to freely ask the researcher questions before data collection began.

As the final step of onboarding, participants put on the wearable devices in the following order: Polar H10, Empatica E4, Muse S, and a headset. The researcher helped adjust the webcam angle and the headset and ensured that everything was functioning properly. Customer-role participants were required to demonstrate their professional acting experience (e.g., a certificate of acting) upon arrival at the site. Then, they read the study

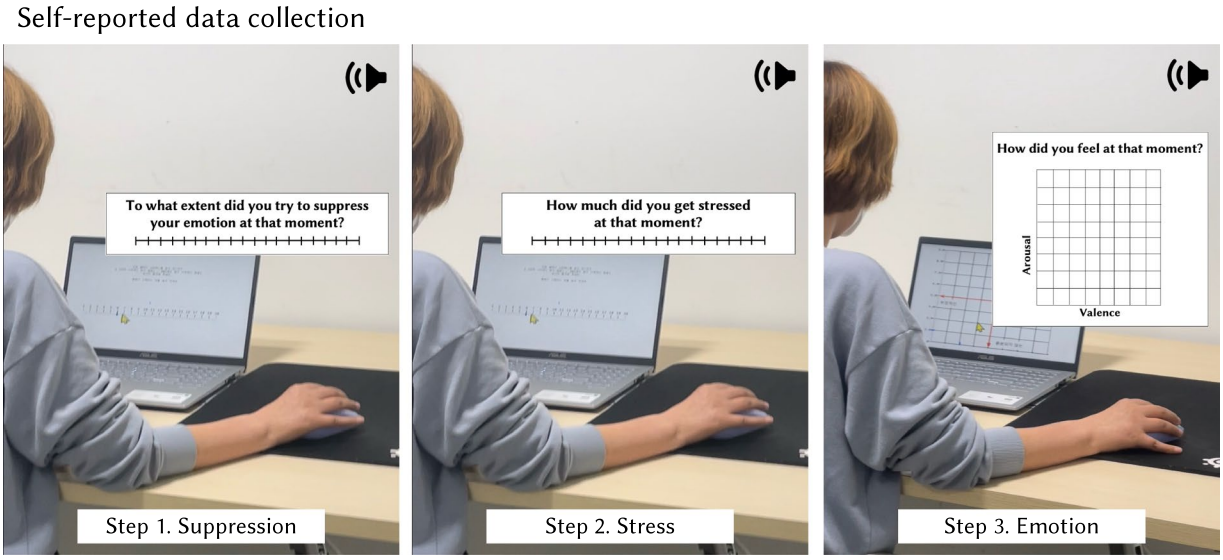


Fig. 2 Label acquisition process. After scenario-based sensor data collection, call agent participants performed self-reported labeling in the order of suppression (i.e., emotional workload), stress, and emotion (i.e., arousal and valence) through retrospective listening to the calls. They listened to the same calls three times in total, once for each labeling process. While listening to the calls, they moved the mouse cursor on the guided images to indicate their state. We used a slider with an anchor value ranging from 1 to 20 (20-point Likert scale) that moved with the cursor for suppression and stress labeling, and utilized an image of a two-dimensional space, which represents arousal and valence along each axis with values ranging from 1 to 9 (9-Point Likert scale) for emotion labeling. The order of calls was identical to that in the sensor data collection.

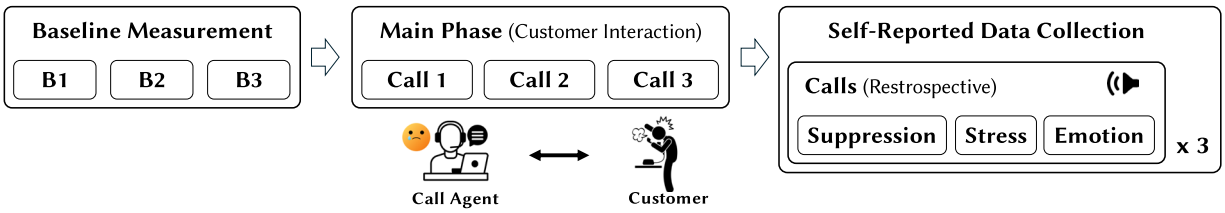


Fig. 3 Overview of the data collection procedure. The experiment consisted of three phases: (1) baseline measurements (B1-B3), (2) the main phase involving three customer interaction calls (C1-C3) under varied emotional conditions, and (3) retrospective self-reported data collection for suppression, stress, and emotion.

instructions and completed the user consent form. A researcher in the designated space explained their role and provided supplementary documents. The documents included fictional character information (e.g., name, age, and phone number) and detailed descriptions of the customer's problem, which they were instructed to act out. They were granted sufficient time to acclimate to the role.

Baseline measurement. Three types of baseline conditions were designed to reflect typical tasks of call agents, such as speaking and typing. Baseline 1 (B1) was a condition with no stimuli, so the participants sat comfortably and were relaxed. Baseline 2 (B2) was a condition in which participants read a neutral script in a natural speaking tone, unrelated to customer service work. B2 (speaking without emotional workload) was designed to compare recorded audio from the main experimental setting, thereby enabling a more in-depth analysis of speech data. Baseline 3 (B3) was a condition in which participants typed a given script. B3 (typing without emotional stress) was designed to isolate the hand movement patterns that arise solely from typing behavior. It has the potential to be used later to normalize or subtract typing-related acceleration signals from the main experimental data, thereby enabling more accurate detection of motion artifact induced by mental status changes, such as stress level or emotion. Each session of these baseline conditions was designed to last approximately two minutes.

Main phase. We simulated a representative example of an abusive customer complaint call in a real call center scenario, where a customer complained about a card company's reward[35,47–49]. Credit card customers who had not received cash rewards contacted the call center to raise an issue due to a misunderstanding of the complex rewards rules. Each call agent answered three calls from the same customer. While the problem remained the same across all calls, the way the customer expressed their complaint differed. Call 1 (C1) represented a neutral scenario. Customers simply explained their problems and asked for the reason behind the problem without

Survey	Measurement	# of items & response scale	Korean adaptation
PHQ-9 (Patient Health Questionnaire-9)[40]	Depressive symptoms	9 items; 4-point Likert	Park *et al.*[76]
PSS (Perceived Stress Scale)[41]	Perceived stress	10 items; 5-point Likert	Lee *et al.*[77]
PPC (Positive Psychological Capital Questionnaire)[42]	Self-efficacy, hope, optimism, and resilience	18 items; 6-point Likert	Lim[78]
BFI (Big Five Inventory)	Five Personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism)	44 items; 5-point Likert	Kim *et al.*[46]
RSE (Rosenberg Self-Esteem Scale)[44]	Global self-esteem	10 items; 4-point Likert	Lee *et al.*[79]
SWLS (Satisfaction with Life Scale)[45]	Cognitive life satisfaction	5 items; 7-point Likert	Lim[80]

Table 2. Summary of psychological survey instruments used in the study and their validated Korean adaptations. All questionnaires used in this study were based on validated Korean translations of the original instruments. In particular, personality traits were measured using the 15-item Korean version of the Big Five Inventory (K-BFI-15).

using emotionally charged language. Call 2 (C2) represented an aggressive scenario involving prosodic stimulation (e.g., talking fast, shouting, and a threatening tone). Note that customers were instructed to avoid using negative words (e.g., swearing) to minimize confusion with C3. Call 3 (C3) represented an aggressive scenario involving expressive stimulation (e.g., demeaning and negative words, including swearing). We placed restrictions on the use of excessively profane expressions and provided a list of permissible expressions to maintain ethical standards. Note that customers were instructed to refrain from prosodic features (e.g., shouting) to minimize confusion with C2.

We defined C1 as less emotional stimuli or basic workload condition, C2 as high emotional prosodic stimuli or high workload condition, and C3 as high emotional expressive stimuli or high workload condition. Each session lasted approximately three to four minutes. The order of the conditions was randomized as C1-C2-C3 or C1-C3-C2 to avoid order effects. C1 was performed first in all cases to allow for a natural flow, while the order of the two different types of aggressive calls (C2 and C3) was varied to prevent order effects. In each call, customers initiated and terminated the conversation while call agents were not allowed to control the start and end of a call, just like in natural inbound calls. Between calls, call agents took a break with a one-minute guided breathing video to release emotional stress caused by the previous call. After completing three calls, the call agents removed the wearable devices and engaged in a 10-minute guided meditation.

Label acquisition. After the meditation session, call agents moved from the sensor data collection room to the self-report data collection room. Following prior work on retrospective continuous self-report during stimulus playback[35], each call was played back in the original order. For each call, participants performed the self-report three times in a row to assess suppression, stress, and emotion (i.e., arousal and valence), as illustrated in Fig. 2. We intentionally chose this post-interaction, playback-based procedure rather than soliciting ratings during or immediately after each call to preserve workflow continuity and ecological validity across the three consecutive interactions and avoid dual-task interference that could alter natural conversational dynamics. Before the assessments began, all participants completed a practice session to familiarize themselves with the interface and the data recording process. Then, they were asked to read the instructions containing the meaning and examples of each label as defined in Table 3. For example, participants listened to the C1 session a total of three times, rating suppression during the first listening, stress during the second, and emotion during the last.

Simulation fidelity. The experiment was conducted in a lab environment designed to reproduce realistic call-center scenarios as closely as possible. Because the study took place under controlled laboratory conditions, participants were informed during the consent process about the general purpose of the study, the number of calls to be conducted, and the types of data to be collected. Although this awareness could potentially influence their behavioral or affective responses, we made extensive efforts to ensure that the interactions resembled authentic customer service situations. To enhance realism, participants were asked to perform their genuine professional roles, and professional actors were recruited to portray customers. Importantly, the call agents were not informed that the customers were professional actors or about the specific behavioral strategies the actors would employ. After data collection, a brief post-experiment questionnaire assessed the perceived reality of the scenario.

Preprocessing and Feature Extraction. We used Python libraries to verify the dataset and then process it to make it suitable for public release.

Labels. The initial results from E-Prime 3.0 were saved in a '.txt' file, including recorded values with timestamps in milliseconds (ms) starting at 0 from the beginning of the call. We checked interval consistency and total duration against the corresponding audio files to ensure temporal integrity, then converted the relative timestamps to absolute Coordinated Universal Time (UTC) for multimodal alignment (details in the Technical Validation section). We added the UTC timestamp of the start of each call to the original E-Prime 3.0 timestamps. Then we saved all transformed data in a dictionary structure and stored it as a pickle file (label.pkl).

Measurement	Questions	Metric	Reference
Suppression	To what extent did you try to suppress your emotion at that moment?	20-point Likert scale	Gabriel and Diefendorff[35]
Stress	How much did you get stressed at that moment?	20-point Likert scale	Gabriel and Diefendorff[35]
Arousal	How do you feel at the moment? (Arousal means 'immediate feeling of excitement vs. calmness')	9-point Likert scale	Russell[75]
Valence	How did you feel at that moment? (Valence means 'immediate feeling of negative vs positive')	9-point Likert scale	Russell[75]

Table 3. Questions and rating scales for acquiring self-reported labels.

Sensors_Muse S, Polar H10, Empatica E4. We release the raw data collected from each device. All details of the collected data are described in Data Records.

Sensor_Audio. We prepared three separate audio recordings for ease of data usage: customer-only, worker-only, and dyadic conversation recordings. We used the customer-only and the worker-only records that were stored at 44,100 Hz in '.wav' format. Because of privacy concerns, audio files are not released to the public. Instead, we extracted well-known features that are widely used in previous studies on emotion and stress recognition, including a deep learning-based approach[50]. We used Librosa 10.0 and applied a window size of 25 ms with a sliding window of 10 ms[50]. Before feature extraction, we applied a pre-emphasis filter provided by Librosa. We then extracted features, including fundamental frequency (F0), zero crossing rate (ZCR), voiced flag, mel-frequency cepstral coefficients (MFCCs), and spectral features (centroid, bandwidth, roll-off, roll-off min, contrast)[51,52] All transformed data are saved in a dictionary structure and stored as a pickle file (audio.pkl). The dyadic conversation recordings were used for script generation to help other researchers gain a comprehensive understanding of our scenario. Because the participants were Korean, all conversations were recorded in Korean. Thus, we used the Naver Clova Speech-to-Text (STT) service, which is optimized for Korean. A researcher verified the accuracy of each transcript and manually modified the text in Korean. The researcher then automatically translated it into English using ChatGPT-4o, followed by manually verifying the translated script. To balance accessibility with linguistic fidelity, we release the original Korean utterances alongside their English translations. This follows established practice in non-English speech corpora that preserve source-language transcripts to support downstream research (e.g., cross-lingual modeling, and prosody-text analyses)[53,54]. This data was transformed into JavaScript Object Notation Lines (JSONL) files at the utterance level.

Sensor_Video. We recorded the upper body of the call agents as an '.mp4' file, which was recorded at 30 frames per second (FPS). From the videos, we extracted features relevant to emotion and stress detection (i.e., head movements, lip and eye movements) from face landmarks[55,56] utilizing OpenCV and MediaPipe. We extracted Mediapipe face_detection and face_mesh functions (confidence = 0.7) to extract 468 landmarks on the face. From each frame, the following features were extracted based on facial landmarks. If a frame was not successfully read or facial landmarks could not be detected, feature extraction was not performed. Nevertheless, a timestamp was generated for every attempted frame regardless of detection success. An error_type column was introduced to indicate the outcome of each detection attempt, with values defined as follows: 0 for success, 1 for failure to read the frame, and 2 for failure to detect the face and landmarks. We obtained rotation and translation vectors (3×1 matrices) from detected landmark points, as well as the total translation computed from the translation vectors. We determined eye-blink events between frames using the Eye Aspect Ratio (EAR), calculated from six landmarks around each eye. To estimate eye gaze direction, we extracted data from four iris landmarks per eye with reference to the facial center. For lip corner movements, we computed relative positions of a landmark at the left and right corners of the mouth with respect to the facial center (i.e., a point of the nose tip). All extracted data were stored in a dictionary structure and saved as a pickle file (video.pkl).

Data Records

The dataset is available at EmoWork[57]. We collected a multimodal dataset from each session (Table 4). A total of 31 call agents (24 female, 7 male) data were collected, and the information about the participants is summarized in Table 5. EmoWork consists of subfolders: META, SENSORS, LABELS, and CSV.

META. The META folder includes helpful information to understand the collected dataset and facilitate further analysis, which has the following files:

- meta.csv: Pnum (subject number), Order_type (0: C1-C2-C3, 1: C1-C3-C2), Actor (subject number of customer), Actor_sex (M: Male, F: Female).
- pre_survey.csv: Pnum (subject number), Sex (0: Male, 1: Female), Age, Education level (0: high school, 1: above bachelors degree), Dominant hand (0: right, 1: left), Career (as a call agent, years), and responds to each questionnaire for PHQ-9 (nine questions), PSS (ten questions), PPC (18 questions), BFI-15 (15 questions), RSE (ten questions), and SWLS (five questions).
- post_survey.csv: Pnum (subject number), Reality of scenario (0: no, 1: yes), Reality of scenario (7-point Likert scale), B1_relaxed (0: no, 1: yes), B2_relaxed (0: no, 1: yes), B3_relaxed (0: no, 1: yes), Stress in C1 (7-point

Role	Device	Sampling Rate	Data	B1	B2	B3	C1	C2	C3
Customer, Worker	Jabra Biz 2300 USB MS Mono	44,100 Hz	Audio derived features with a time-window size of 25 ms and a 10 ms sliding window (F0, Energy, ZCR, voiced-flag, MFCCs, spectral centroid, spectral bandwidth, roll-off, roll-off min, spectral contrast)	x	o	x	o	o	o
Worker	Logitech C270	30 FPS	Video derived features (head movements, lip movement, and eye movements)	x	x	x	o	o	o
Worker	Muse S	256 Hz	EEG (TP9, AF7, AF8, TP10)	o	o	o	o	o	o
Worker	Polar H10	130 Hz	ECG, heart rate, RR-intervals	o	o	o	o	o	o
Worker	Empatica E4	BVP (64 Hz) EDA (4 Hz) Temp (4 Hz) ACC (32 Hz)	BVP, EDA, Temperature, and ACC (x,y, and z)	o	o	o	o	o	o
Worker	E-Prime 3.0	5–10 Hz	Self-reported labels (stress, arousal, valence, and suppression)	x	x	x	o	o	o

Table 4. Summary of dataset. Audio and video data are provided as derived features, while other sensor data are shared in their raw form as originally collected from each device. During baseline measurements, audio data were recorded only from the worker in B2. Video data and self-reported labels were only collected from C1, C2, and C3. E-prime 3.0 is software installed on a laptop.

Likert scale), Stress in C2 (7-point Likert scale), Stress in C3 (7-point Likert scale), Stress from environmental setting (7-point Likert scale), and Stress from wearable device (7-point Likert scale).

- B2_script_korean.txt: The Korean script used for baseline 2 measurement.
- B2_script_english.txt: The English script translated from the Korean script.
- call_script.jsonl: A dyadic script in both Korean and English transcribed from the STT output call audio. Each line follows data schema of 'participant_id' (i.e., pnum), 'scenario' (i.e., condition), 'speaker', 'turn', 'speaker_turn', 'script_kor', 'script_eng'. Only the value of turn and speaker_turn are saved as integers, and others as strings.

LABELS. The LABELS folder contains labels.pkl, which include momentary stress, arousal, valence, and emotional workload (i.e., suppression) during each call. The file follows a hierarchical structure where each participant (i.e., pnum) contains conditions (C1, C2, and C3), and each condition stores labels (stress, arousal, valence, suppression) as keys, and with self-reported data with a timestamp as values stored as a dataframe. The mean interval between perceived stress records was 101.6 ms (SD = 7.5, Min = 100, Max = 381) ; suppression was 101.6 ms (SD = 7.7, Min = 100, Max = 331) ; arousal and valence were 102.3 ms (SD = 6.5, Min = 101, Max = 368).

SENSORS. The SENSORS folder contains audio.pkl, sensor.pkl, and video.pkl. All files have the same hierarchical dictionary structure, where each participant (pnum) contains different conditions (B1, B2, B3, C1, C2, and C3). Table 6 describes the details of all extracted features.

Sensors. The sensor.pkl contains data from Muse S, Empatica E4, Polar H10, and Galaxy PPG. Under each condition (e.g., C1), there are subdictionaries with keys from muse, e4_acc, e4_bvp, e4_ibi, e4_eda, e4_temp, Polar_ecg, Polar_hr, Polar_rr, Galaxy_ppg. Under each sensor as a key, there is data with a timestamp as a dataframe. The codes to check each data included in the supplied code.

- Muse S: Muse contained four electroencephalogram (EEG) signals from temporal posterior (TP) 9, anterior frontal (AF) 7, AF8, and TP10 in the international EEG 10-10 system placement with a timestamp. The total amount of expected muse data was 8,877,167, and the actual data points are 8,750,751 (1.42% difference). The actual data points include 256 duplicate timestamp data points with different values of EEG data. 27 among 186 sessions (six conditions from 31 participants) had abnormal intervals (e.g., over 10 milliseconds); four participants in B1 (P1, P9, P18, P30), six in B2 (P1, P10, P15, P17, P22, P25), five in B3 (P9, P17, P19, P22, P29), three in C1 (P1, P16, P17), four in C2 (P12, P21, P26, P30), and five in C3 (P4, P8, P9, P14, P26). Also, EEG was recorded in C1 of P13, with a notably shorter duration (95.57 seconds) compared to other sensors, which typically recorded over 200 seconds.
- Empatica E4: As each sensor has a different interval, data from the Empatica E4 were saved separately in 'e4_acc,' 'e4_eda,' 'e4_temp,' 'e4_bvp,' and 'e4_ibi'. 'e4_acc' contained accelerometer (ACC) data from the x, y, and z axes with a timestamp from the ACC sensor, 'e4_eda' contained electrodermal activity (EDA) with a timestamp from the EDA sensor, and 'e4_temp' contained temperature (Temp) with a timestamp from the optical thermometer sensor. 'e4_bvp' contained a blood volume pulse (BVP) with a timestamp, and 'e4_ibi' contained an interbeat interval with a timestamp from a PPG sensor. We had one session which does not have all the data from E4 (B2 of pnum 25). We examined their intervals to confirm the amount of data based on its frequency. The difference between the expected data point and the actual data point was less than 0.1% in 'e4_acc,' 'e4_eda,' 'e4_temp,' and 'e4_bvp'. However, 'e4_ibi', the calculated heartbeat interval (ms) provided by the device from the BVP signal, was not collected well. According to the company's guidelines, IBI data can be extracted assuming a clean BVP signal. Therefore, we include BVP data and IBI data collected from PPG signals in our public data for future researchers to utilize, but exclude them from technical validation.

Category	Distribution
Sex	Female = 24, Male = 7
Education	Graduated from high school = 19, Graduated from college or university = 12
Dominant hand	Right = 30, left = 1
Age (year)	Average = 36.19 (SD = 11.38)
Career (year)	Average = 3.29 (SD = 2.49)
PHQ-9	No depression = 19, Light depression = 9, Moderate depression = 3, Moderately severe depression = 0, Severe depression = 0
PSS	Low stress = 10, Moderate stress = 18, High stress = 3
RSE	Low self-esteem = 4, Normal = 22, Above normal = 5
SWLS	Extremely dissatisfied = 1, Dissatisfied = 7, Below average = 8, Average = 7, Highly satisfied = 7, Very highly satisfied = 1

Table 5. Demography information of participants. This table presents the demographic distribution of participants, including sex, education level, dominant hand, age, and career length as call agents. Additionally, it summarizes psychological measures such as depression (PHQ-9), perceived stress (PSS), self-esteem (RSE), and life satisfaction (SWLS). All questionnaires for each measurement and information for calculation is included in shared code.

- Polar H10: Due to the difference in the intervals of records, we saved data from the Polar H10 separately in 'Polar_ecg', 'Polar_hr', and 'Polar_rr'. 'Polar_ecg' contains ECG data with a timestamp, 'Polar_hr' contains HR data with a timestamp, and 'Polar_rr' contains RR interval data with a timestamp.

Audio. The audio.pkl contains the prosodic features extracted from customer-only.wav and worker-only.wav. For portability, a CSV mirror of .pkl is provided and documented in. No raw audio is included in the public release. Baselines were collected from the worker only in 'b2'. 'c1,' 'c2,' and 'c3' store the data in a dictionary structure with separate key values for workers and customers. The dictionary uses the numeric keys: 1 (worker) and 2 (customer), each of which maps to a dataframe consisting of timestamps, F0, energy, ZCR, voiced flag, MFCCs (1-13), and spectral features (i.e., centroid, bandwidth, roll-off, minimum roll-off, and contrast). Since all participants had complete audio recordings, no missing data were found. However, for P1 and P2, the baseline was measured for only one minute, resulting in approximately one minute of data, whereas for all other participants, it was measured for two minutes.

Video. The video.pickle and video folder in CSV folder contain extracted features data from video.mp4 files. The extracted video data were stored in a subject- and condition-specific dictionary structure, consistent with other sensor data. Each key of 'c1,' 'c2,' 'c3' has the value of a data frame including timestamp, rotation_x, rotation_y, rotation_z, translation_x, translation_y, translation_z, total translation, blink, gaze_x, gaze_y, left lip corner movements, right lip corner movements, and error type. A total of 395 minutes of video were collected from 31 participants. The expected number of data points was 711,000 (395 x 60 s x 30 FPS), and the actual number of data points was 710,738 (difference: 0.04%). Among the collected data points, 1.88% of the data points had a missing value due to error type 2 (failure to detect faces and landmarks). Individuals had an average missing value of 1.75%. Nine participants with a higher than average percentage of missing data exhibited behaviors that interfered with automatic detection. It is inferred that certain behaviors caused the failure to detect faces or landmarks, such as wearing glasses that reflected light (P25, P29, P2, P19); ducking the face (P5, P6, P7), covering the mouth with a hand or holding the microphone (P20, P22), or being so close to the screen that only part of the face was captured in the image (P7).

CSV. To improve portability and random access, we also decomposed each original pickle file into a folder with files of '.csv' layout that preserves the original hierarchy. As described in Fig. 4, each key in the original pickle file corresponds to a folder inside the CSV folder. All final-level dataframes are stored as individual '.csv' files in the original pickle hierarchy. For example, LABEL[1]['c1']['stress'], previously a pandas dataframe, is now stored as LABEL/1/c1/stress.csv.

- LABEL_csv (from label.pkl): Multiple nested keys in the original structure (participant → condition → labels) are provided separately as each csv file per participant, condition (i.e., C1, C2, C3), and label type. The results yield a total of 372 files ($31 participants \times 3 scenario \times 4 labels = 372$).
- VIDEO_csv (from video.pkl): Each participant has 3 conditions (i.e., C1, C2, C3), resulting in 93 files in total ($31 participants \times 3 conditions = 93$).
- SENSOR_csv (from sensors.pkl): For 19 among 31 participants, each has data from 6 scenarios (i.e., B1, B2, B3, C1, C2, C3) and 10 sensor types. For 12 among 31, data include 9 sensor types per scenario; one is missing from Galaxy PPG. This results in a total of 1,788 CSV files ($(19 participants \times 6 conditions \times 10 sensors) + (12 participants \times 6 conditions \times 9 sensors) = 1,788$).
- AUDIO_csv (from audio.pkl): Each participant has 3 conditions (i.e., C1, C2, C3) with 2 audio channels (i.e., worker and customer). Additionally, there is one more record per worker in one of the baseline conditions (i.e., B2). In total, this results in 217 files ($(31 participants \times 3 conditions \times 2 channels) + (31 participants \times 1 conditions \times 1 channels) = 217$).

Sensor	Data	Unit	Definition
Audio	F0	Hz	Fundamental frequency (63 Hz–2,093 Hz, librosa.pyin), null = 0
Audio	Energy	—	Root mean square of energy (librosa.feature.rms)
Audio	ZCR	—	Rate of zero crossings in the signal (librosa.feature.zero_crossing_rate)
Audio	Voiced Flag	0 or 1	Binary indicator for detected fundamental frequency (63 Hz-2,093 Hz, librosa.pyin)
Audio	MFCCs 1–13	—	13 mel-frequency cepstral coefficients representing timbre (librosa.feature.mfcc, Hamming window applied)
Audio	Spectral Centroid	Hz	Center of mass of the spectrum (librosa.feature.spectral_centroid, Hamming window applied)
Audio	Spectral Bandwidth	Hz	Spread of frequencies around the spectral centroid (librosa.feature.spectral_ bandwidth, Hamming window applied)
Audio	Roll-off	Hz	Frequency below which a 99% of the total signal energy is contained (librosa.feature. spectral_rolloff, roll percent = 0.99)
Audio	Roll-off min	Hz	Frequency below which a 1% of the total signal energy is contained (librosa.feature. spectral_rolloff, roll percent = 0.01)
Audio	Spectral contrast	—	Difference between spectral peaks and valleys (librosa.feature.spectral_contrast)
Video	Rotation_x, y, z	radian	3 Rodrigues rotation vectors from head movemenets (cv2.solvePnP)
Video	Translation vector_x, y, z	—	3D translation (or shift) between the world coordinate system and the camera coordinate system (cv.solvePnp)
Video	Translation vector_total	—	Absolute distance of face from camera
Video	Blink	—	A binary flag indicating a change in EAR (Eye Aspect Ratio). If one of the eyes' EAR was below 0.3 in the previous frame and is $\geq$ 0.3 in the current frame, set to 1; otherwise, 0.
Video	Gaze X	—	Relative horizontal position of the iris center with respect to the nose tip (nose tip: landmark 1, left iris: 474, 475, 476, 477, right iris: 469, 470, 471, 472)
Video	Gaze Y	—	Relative vertical position of the iris center with respect to the nose tip (nose tip: landmark 1, left iris: 474, 475, 476, 477, right iris: 469, 470, 471, 472)
Video	Left Lip Corner	—	Absolute distance of lip corner from the nose tip (nose tip: landmark 1, left lip corner: 61)
Video	Right Lip Corner	—	Absolute distance of lip corner from the nose tip (nose tip: landmark 1, right lip corner: 291)
Video	Error type	—	Type of feature extraction results (0 : success, 1: fail to read frame, 2: fail to detect face and landmarks)
Muse S	TP9	μV	Electroencephalogram (EEG) signals from temporal posterior on the left side
Muse S	AF7	μV	EEG signals from anterior frontal on the left side
Muse S	AF8	μV	EEG signals from anterior frontal on the right side
Muse S	TP10	μV	EEG signals from temporal posterior on the right side
Polar H10	ECG	μV	Electrical activity of the heart, captured using electrocardiography
Polar H10	Heart Rate (HR)	BPM	Number of heartbeats per minute
Polar H10	RR Intervals	ms	Time interval between consecutive R-wave peaks in the ECG signal
E4	EDA	μS	Electrodermal activity, measuring skin conductance as an indicator of physiological arousal
E4	BVP	—	Blood volume pulse, measuring cardiovascular activity
E4	IBI	ms	Interbeat interval, the time between successive heartbeats
E4	Skin Temperature	°C	Temperature measured at the skin surface
E4	Acceleration	m/s^2	3-axis accelerometer data capturing wrist movement

Table 6. Types of sensor and data definition. This table provides information on the collected data, including sensor type, data type, measurement units, and feature definitions. For extracted features (e.g., audio and video), details of the feature extraction functions are also provided.

Technical Validation

Simulation Fidelity Assessment. We assessed ecological validity with evidence that the laboratory scenario resembles real call-center work using the post survey on simulated scenarios. After each session, agents completed a brief post-experiment questionnaire including (1) a binary realism item and (2) a 7-point similarity item (1 = not at all similar, 7 = very similar) regarding resemblance to real work. On the binary item, 29 of 31 call-center agents (93.5%) indicated that the scenarios were realistic. On the 7-point item, agents rated the scenarios as similar to real work (M = 5.90, SD = 1.40). These results suggest a reasonable level of simulation fidelity.

Self-Reported Scale Distribution Analysis. We verified that the experimental manipulation (i.e., normal complaint, shouting, swearing) was reflected in self-reported affective measurements, supporting the validity of the condition and any derived labels for modeling. Moreover, as the other data descriptor presented the information on annotated data[58], these analyses provide an intuitive understanding of perceived metrics. Therefore, we summarized condition-wise distributions for stress, emotional workload, valence (1 = negative, 9 = positive), and arousal (1 = calm, 9 = aroused). Figure 5 shows the distributions and episode-level traces. Stress increased across conditions (normal: M = 4.29, SD = 4.19; swearing: M = 10.86, SD = 6.38; shouting: M = 14.08, SD =

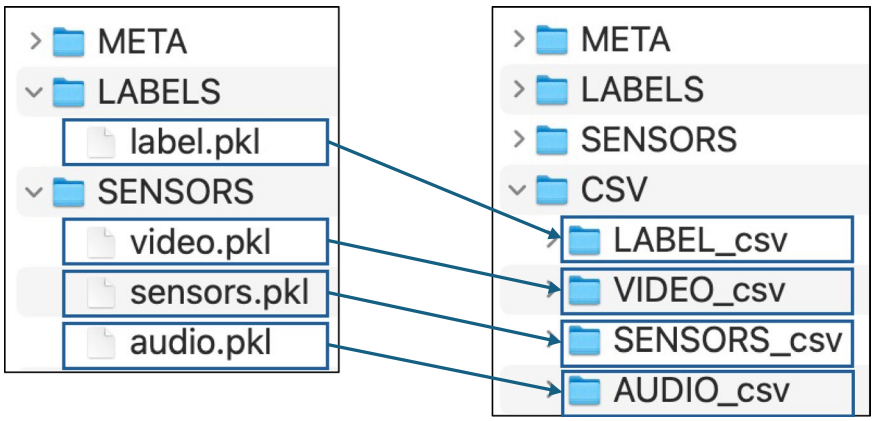


Fig. 4 Relationship between the original pickle (.pkl) files and the CSV folder layout. Each .pkl is decomposed into a portable folder (CSV hierarchy: dictionary keys map to subfolders, and each leaf-level dataframe is saved as an individual csv. Arrows indicate the mapping).

6.05). Emotional workload showed a similar increase (normal: M = 5.29, SD = 5.13; swearing: M = 11.03, SD = 6.43; shouting: M = 13.01, SD = 6.51). Valence decreased (normal: M = 4.63, SD = 1.69; swearing: M = 2.86, SD = 1.60; shouting: M = 2.45, SD = 1.58), while arousal increased (normal: M = 4.39, SD = 1.76; swearing: M = 5.25, SD = 2.07; shouting: M = 5.86, SD = 2.25). The traces demonstrate momentary variation within episodes consistent with the manipulation.

Time-synchronization. We collected multimodal data from multiple devices; thus, it should be ensured that multimodal streams recorded on different devices are time-aligned for valid cross-modal data analyses[28]. Devices were synchronized to UTC, but the time to start recording was inconsistent. Therefore, for each participant and condition, we verified that device streams shared a common time range, defined as the intersection of their start and end timestamps. We define the overlap ratio as the duration of the common time interval divided by the minimum device-specific interval within a session. In results, across phone-call intervals, the integrated data duration averaged 225.46 seconds (SD = 21.36, min = 95.57, max = 273.54), indicating consistent session lengths appropriate for multimodal alignment, with verified overlap across devices for all sessions.

Dataset Validation using Machine Learning Models. As many recent data descriptors use the machine learning models and deep learning models to validate the utility of the data records[27,31,59], we trained machine learning models to classify (1) workload intensity based on interaction condition and (2) perceived affect states from multimodal signals. Here, we aim to demonstrate the presence of learnable signals for checking data quality and reusability, not a benchmarking or hypothesis-testing exercise.

We formulated two tasks. 1) *Condition classification*: normal conversations were labeled as low-intensity workload (0) and shouting/swearing as high-intensity (1); 2) *Perceived affect classification*: we derived binary target labels for perceived stress, emotional workload (i.e., suppression), arousal, and valence from self-reported ratings via participant-specific thresholds (per-variable mean). We used audio-derived features; E4 temperature, EDA, and ACC; Polar H10 ECG; and Muse S EEG. Preprocessing (normalization, filtering, feature extraction) was *fitted on training folds only* and applied to the held-out fold to avoid leakage; processed multimodal features were then integrated for modeling. Classifiers included Random Forest (RF)[60], eXtreme Gradient Boosting (XGBoost)[61], Decision Tree (DT)[62], Support Vector Machine (SVM)[63], and k-Nearest Neighbors (kNN)[64], commonly used in affective computing[27,28,65]. Evaluation used Leave-One-Subject-Out (LOSO; $N = 31$) to estimate subject-independent generalization.

Summary metrics are reported in Table 7 for task 1 and Table 8 for task 2. We report the mean of AUC, accuracy, F1-score, Precision, and Recall with standard deviation across held-out participants. Figure 6 shows per-participant performance by model for task 1. RF achieved a mean AUC of 0.946 across LOSO folds, with lower values concentrated in a small subset of participants (IDs 25-29); a similar pattern was observed for other algorithms (Table 7). In our protocol, each participant interacted with one of four actors (e.g., P25-P27 with Actor 3; P28-P29 with Actor 4). Because a single actor delivered all conditions for a given participant, part of the between-subject variance likely reflects actor-specific delivery as well as personal differences. Users building general models may consider including actor identity or conducting actor-robust validation. For task 2, participants with single-class outcomes due to highly skewed ratings were excluded (stress: P22, P28; suppression: P28, P30; valence: P1).

Usage Notes

The archive is released under a controlled-access policy, because several data components (i.e., high-resolution physiological signals, obfuscated voice/video features) contain *indirect identifiers* and other *sensitive fields*. Researchers wishing to use the data should carefully read the guidelines in Zenodo, then 1) download the `EmoWork_DUA.pdf`, 2) complete the data-usage agreement that specifies research purpose, storage safeguards, and a commitment to non-re-identification, and 3) submit a request on Zenodo.

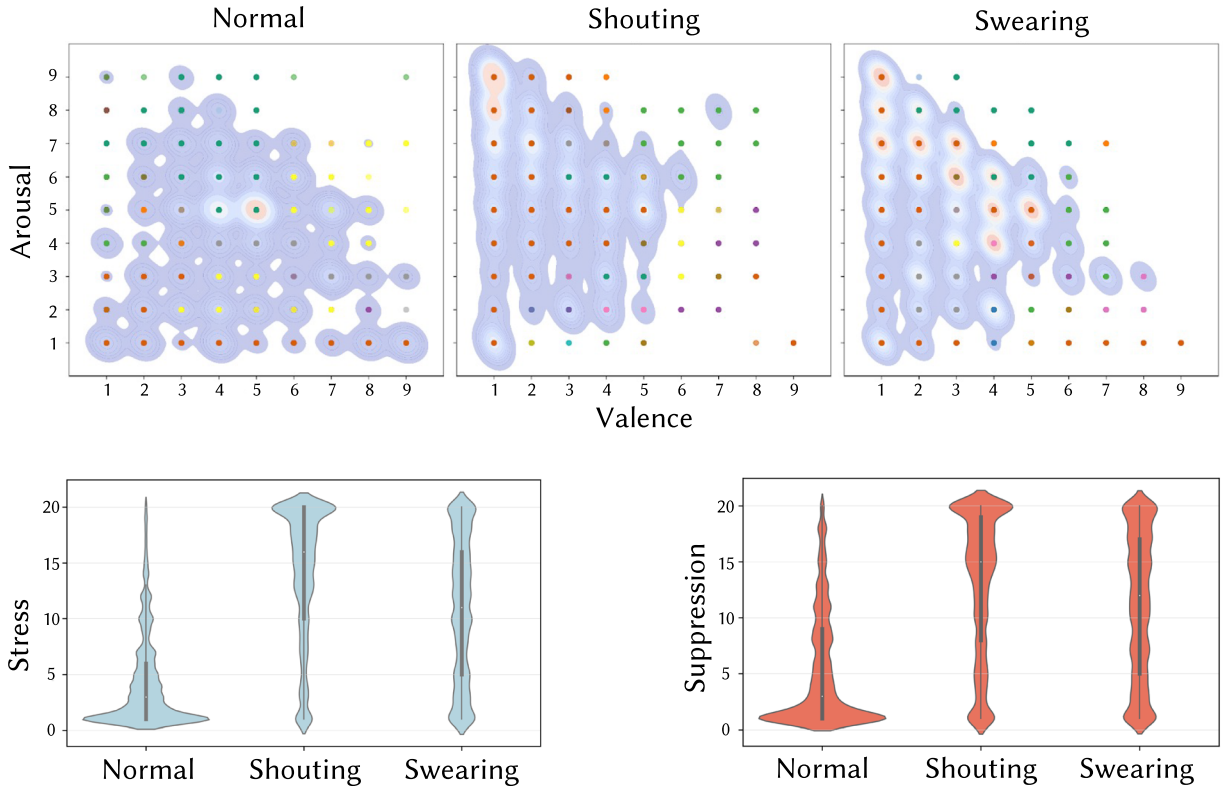


Fig. 5 Distributions of self-reported labels across experimental conditions. The top row shows kernel density plots of arousal and valence ratings in each condition (Normal, Shouting, Swearing). The bottom row illustrates violin plots of perceived stress (left) and suppression (right) for each condition. Compared to the Normal condition, Shouting and Swearing conditions elicited higher stress, suppression, and arousal with lower valence.

Handling Incomplete Audio and Video Segments. *Audio baseline.* The baseline assessment before conversation was designed to last 2 min; however, two of the 31 participants recorded only the first 60 s. We suggest three possible handling strategies. First, users may use a standardized 60 s baseline, because all recordings share the same temporal anchor (start at t = 0). Second, users can just use data without a baseline and focus on relative changes between conditions. Lastly, the other possible strategy is excluding these two participants (strict subset). Users should report the selected option and any related sensitivity analyses.

Video features. We provide an error_type flag that distinguishes no error, face-detection failure, and feature-extraction failure. By default, we suggest discarding frames flagged as `error_typesuccess` (0) and dropping windows whose missing-data rate exceeds a defined threshold. For continuous kinematic features (e.g., gaze or lip-corner movement), interpolation or smoothing (e.g., Kalman filtering[66]) can be applied, while imputation of discrete outputs is discouraged. Users should report the selected option and any related sensitivity analyses.

Possible Applications. This multimodal dataset was collected in a specific work scenario involving call agents under emotional workload, requiring them to suppress negative emotions and express positive emotions during voice-based customer interactions. It includes a package of labels measured by Likert scale (i.e., emotion, stress, and emotional workload) suited for a specific context, along with multimodal data readily obtainable in everyday environments. Therefore, in the call-center domain, the data can be used in realizing many applications, such as employee training systems with simulated customers to help new employees practice emotional regulation in difficult calls; operational monitoring to identify high-workload episodes during voice-only customer service; and just-in-time interventions that utilize the results of automatic assessments for coaching or enhancing workers' well-being. With appropriate domain adaptation, the dataset can inform real-time psychological state monitoring to adjust task difficulty in mission-critical teams training (e.g., flight or space crews[67] and military training[68]) that involve scenario dynamics under high mental workload and voice-based communication. It may also extend to other public-facing professions that routinely manage emotionally demanding interactions (e.g., flight attendants and frontline public-service staff). The dataset can also be used to develop toolkits (e.g., Han *et al.*[69]) for assessing and validating algorithms that classify perceived stress, emotions, and emotional workload. Furthermore, this dataset can be used for privacy-preserving technologies such as encryption of physiological signals, as well as for educational purposes (e.g., signal data preprocessing example).

Models	AUC (std)	Accuracy (std)	F1-score (std)	Precision (std)	Recall (std)
RandomForest	0.946 (0.110)	0.868 (0.150)	0.891 (0.165)	0.890 (0.122)	0.923 (0.192)
SVM	0.914 (0.100)	0.840 (0.107)	0.875 (0.108)	0.881 (0.081)	0.884 (0.143)
XGBoost	0.903 (0.172)	0.838 (0.163)	0.871 (0.178)	0.858 (0.155)	0.926 (0.197)
kNN	0.770 (0.098)	0.747 (0.073)	0.822 (0.061)	0.773 (0.058)	0.884 (0.093)
DecisionTree	0.751 (0.159)	0.768 (0.145)	0.814 (0.139)	0.837 (0.121)	0.822 (0.191)

Table 7. Technical validation of dataset using five basic machine learning models. The reported values are the average performance across all test sets, and the standard deviation (std) indicates the variation among the test sets.

Models	Stress (mean)	Stress (std)	Suppress (mean)	Suppress (std)	Valence (mean)	Valence (std)	Arousal (mean)	Arousal (std)
RandomForest	0.783	0.132	0.739	0.115	0.745	0.166	0.649	0.117
SVM	0.731	0.098	0.677	0.102	0.690	0.113	0.612	0.085
XGBoost	0.732	0.165	0.718	0.172	0.713	0.169	0.637	0.173
kNN	0.645	0.105	0.586	0.063	0.609	0.102	0.541	0.064
DecisionTree	0.567	0.141	0.581	0.101	0.560	0.147	0.542	0.081

Table 8. Summary of classification results of perceived stress, suppression, valence and arousal using five basic Machine Learning Models (AUC).

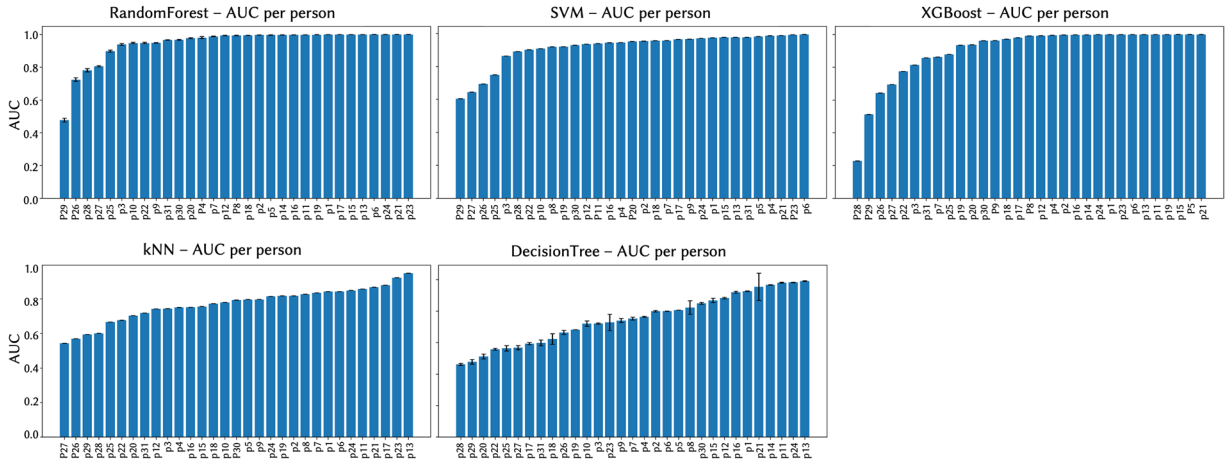


Fig. 6 Area Under the Curve (AUC) scores for each test dataset (i.e., participant) across five machine learning models to classify experimental condition (low emotional workload vs. high emotional workload): Random Forest, Support Vector Machine (SVM), XGBoost, k-Nearest Neighbors (kNN), and Decision Tree. Each bar represents the average AUC over cross-validation folds for a single participant, sorted in ascending order per model. Substantial interpersonal variability was observed in model performance in the data from some participants.

Limitations. This dataset is limited to data collected from a single country (i.e., South Korea), which may not reflect cultural diversity in emotional responses (e.g., body language) that are highly correlated with culture[70,71]. The audio data for prosodic features is also language-dependent data[72]. Therefore, caution is required when applying this dataset to the groups that have different cultural backgrounds. Beyond the cultural aspect, the sample size is relatively small (N = 31), and the sex distribution is imbalanced (7 males). Although such an imbalance is common in the South Korean call-center industry[37], it may still restrict the generalizability of analyses even within the same cultural context, given the potential influence of sex differences in emotional labor[73]. Other demographic characteristics (e.g., age, career, and organization type) are also limited in diversity, which further constrains the dataset's representativeness of the broader call center workforce. Separately, issues of data accessibility and completeness should also be noted. Specifically, the public release contains only extracted features from audio and video data without raw recordings, which may limit the reusability of the data for studies requiring access to the original audio signals. Within the audio modality, two participants' baseline audio segments were shorter than planned (60 s instead of 2 min). Within the video modality, some features are incomplete due to occasional frame or face detection failures because of challenges inherent to naturalistic recording environments (e.g., reflections from glasses, facial occlusion, or movement). For audio and video features, users should consult the Usage Notes for recommended handling procedures. Finally, in the technical validation, the label encoding used in technical validation focused on balancing label distribution, without considering the meaning of arousal and valence itself. Future studies may incorporate these meanings in accordance with their specific objectives.

Data availability

All data supporting this data descriptor are archived on Zenodo at version https://doi.org/10.5281/zenodo.17513978 (Concept DOI: https://doi.org/10.5281/zenodo.15181219), with the current public release v3 under the CC BY-NC-ND 4.0 license. The archive provides participant-level metadata; continuous self-report traces sampled at 100 ms (suppression, stress, arousal, valence) as labels for classification; synchronized features derived from Polar H10, Empatica E4, and Muse S; audio- and video-derived features; and bilingual dyadic transcripts (Korean originals with English translations). Tabular components are available in '.csv' for portability, and a consolidated pickle archive is retained for convenience. Transcripts are provided in '.jsonl'. For privacy reasons (re-identification risk), raw audio and raw video are not shared. Access is controlled; the application workflow and usage restrictions are described in the *Usage Notes*. Please cite the dataset as Ref. [57]; future updates will appear as new versions under the same concept DOI.

Code availability

The code for using this dataset is accessible via GitHub (https://github.com/Kaist-ICLab/EmoWork.git). It includes basic code to explore data records and is to be used for technical validation. In `requirements.txt`, all required libraries (e.g., cvxEDA[74]) are included. All details are documented with step-by-step comments in the repository.

Received: 4 June 2025; Accepted: 24 December 2025;
Published online: 08 January 2026

References

1. Picard, R. W. *Affective computing* (MIT press, 2000).
2. D'mello, S. K. & Kory, J. A review and meta-analysis of multimodal affect detection systems. *ACM Comput. Surv.* **47**, https://doi.org/10.1145/2682899 (2015).
3. Leite, I., Castellano, G., Pereira, A., Martinho, C. & Paiva, A. Empathic robots for long-term interaction: evaluating social presence, engagement and perceived support in children. *International Journal of Social Robotics* **6**, 329–341, https://doi.org/10.1007/s12369-014-0227-1 (2014).
4. Heinisch, J. S. *et al*. Physiological data for affective computing in hri with anthropomorphic service robots: the affect-hri data set. *Scientific Data* **11**, 333, https://doi.org/10.1038/s41597-024-03128-z (2024).
5. Liu, C.-Y. & Yin, B. Affective foundations in ai-human interactions: Insights from evolutionary continuity and interspecies communications. *Computers in Human Behavior* **161**, 108406, https://doi.org/10.1016/j.chb.2024.108406 (2024).
6. Lim, J., Koh, Y., Kim, A. & Lee, U. Exploring context-aware mental health self-tracking using multimodal smart speakers in home environments. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–18, https://doi.org/10.1145/3613904.3642846 (2024).
7. Schaule, F., Johanssen, J. O., Bruegge, B. & Loftness, V. Employing consumer wearables to detect office workers' cognitive load for interruption management. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* **2**, 1–20, https://doi.org/10.1145/3191764 (2018).
8. Züger, M., Müller, S. C., Meyer, A. N. & Fritz, T. 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*, 1–14, https://doi.org/10.1145/3173574.3174165 (2018).
9. Fridman, L., Reimer, B., Mehler, B. & Freeman, W. T. Cognitive load estimation in the wild. In *Proceedings of the 2018 chi conference on human factors in computing systems*, 1–9, https://doi.org/10.1145/3173574.3174226 (2018).
10. Lasi, H., Fettke, P., Kemper, H.-G., Feld, T. & Hoffmann, M. Industry 4.0. *Business & information systems engineering* **6**, 239–242, https://doi.org/10.1007/s12599-014-0334-4 (2014).
11. Park, E., Jung, Y., Kim, I. & Lee, U. 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*, 1–16, https://doi.org/10.1145/3544548.3581457 (2023).
12. Teoh Yi Zhe, I. & Keikhosrokiani, P. Knowledge workers mental workload prediction using optimised elanfis. *Applied Intelligence* **51**, 2406–2430, https://doi.org/10.1007/s10489-020-01928-5 (2021).
13. Alberdi, A., Aztiria, A. & Basarab, A. Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. *Journal of biomedical informatics* **59**, 49–75, https://doi.org/10.1016/j.jbi.2015.11.007 (2016).
14. Jebelli, H., Choi, B. & Lee, S. Application of wearable biosensors to construction sites. i: Assessing workers' stress. *Journal of Construction Engineering and Management* **145**, 04019079, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001729 (2019).
15. Umematsu, T., Sano, A., Taylor, S., Tsujikawa, M. & Picard, R. W. Forecasting stress, mood, and health from daytime physiology in office workers and students. In *2020 42nd annual international conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 5953–5957, https://doi.org/10.1109/EMBC44109.2020.9176706 (IEEE, 2020).
16. Morshed, M. B. *et al*. 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)*, 1–8, https://doi.org/10.1109/ACII55700.2022.9953824 (IEEE, 2022).
17. Tani, Y., Fukuda, S., Matsuda, Y., Inoue, S. & Arakawa, Y. Workersense: Mobile sensing platform for collecting physiological, mental, and environmental state of office workers. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, 1–6, https://doi.org/10.1109/PerComWorkshops48775.2020.9156085 (IEEE, 2020).
18. Yau, J. C. *et al*. Tiles-2019: A longitudinal physiologic and behavioral data set of medical residents in an intensive care unit. *Scientific Data* **9**, 536, https://doi.org/10.1038/s41597-022-01636-4 (2022).
19. Gjoreski, M., Luštrek, M., Gams, M. & Gjoreski, H. Monitoring stress with a wrist device using context. *Journal of biomedical informatics* **73**, 159–170, https://doi.org/10.1016/j.jbi.2017.08.006 (2017).
20. Smets, E. *et al*. Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *NPJ digital medicine* **1**, 67, https://doi.org/10.1038/s41746-018-0074-9 (2018).
21. Koelstra, S. *et al*. Deap: A database for emotion analysis ;using physiological signals. *IEEE Transactions on Affective Computing* **3**, 18–31, https://doi.org/10.1109/T-AFFC.2011.15 (2012).
22. Abadi, M. K. *et al*. Decaf: Meg-based multimodal database for decoding affective physiological responses. *IEEE Transactions on Affective Computing* **6**, 209–222, https://doi.org/10.1109/TAFFC.2015.2392932 (2015).
23. Subramanian, R. *et al*. Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing* **9**, 147–160, https://doi.org/10.1109/TAFFC.2016.2625250 (2018).

24. Katsigiannis, S. & Ramzan, N. Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices. *IEEE Journal of Biomedical and Health Informatics* **22**, 98–107, https://doi.org/10.1109/JBHI.2017.2688239 (2018).
25. Miranda-Correa, J. A., Abadi, M. K., Sebe, N. & Patras, I. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing* **12**, 479–493, https://doi.org/10.1109/TAFFC.2018.2884461 (2021).
26. Saganowski, S. *et al*. Emognition dataset: emotion recognition with self-reports, facial expressions, and physiology using wearables. *Scientific data* **9**, 158, https://doi.org/10.1038/s41597-022-01262-0 (2022).
27. Chen, J. *et al*. A large finer-grained affective computing eeg dataset. *Scientific Data* **10**, 740, https://doi.org/10.1038/s41597-023-02650-w (2023).
28. Schmidt, P., Reiss, A., Duerichen, R., Marberger, C. & Van Laerhoven, K. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, 400–408, https://doi.org/10.1145/3242969.3242985 (2018).
29. Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M. A. & Kraaij, W. The swell knowledge work dataset for stress and user modeling research. In *Proceedings of the 16th International Conference on Multimodal Interaction*, 291-298, https://doi.org/10.1145/2663204.2663257 (Association for Computing Machinery, New York, NY, USA, 2014).
30. Park, C. Y. *et al*. K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data* **7**, 293, https://doi.org/10.1038/s41597-020-00630-y (2020).
31. Lee, M.-H. *et al*. Eav: Eeg-audio-video dataset for emotion recognition in conversational contexts. *Scientific data* **11**, 1026, https://doi.org/10.1038/s41597-024-03838-4 (2024).
32. World Bank. Employment in services (% of total employment) – world development indicators. https://data.worldbank.org/indicator/SL.SRV.EMPL.ZS Accessed 25 May 2025 (2024).
33. Bureau of Labor Statistics, U.S. Department of Labor. Occupational outlook handbook: Customer service representatives. https://www.bls.gov/ooh/ (2025).
34. Hochschild, A. R. Emotion work, feeling rules, and social structure. *American journal of sociology* **85**, 551–575, https://doi.org/10.1086/227049 (1979).
35. Gabriel, A. S. & Diefendorff, J. M. Emotional labor dynamics: A momentary approach. *Academy of management Journal* **58**, 1804–1825, https://doi.org/10.5465/amj.2013.1135 (2015).
36. Zapf, D., Isic, A., Bechtoldt, M. & Blau, P. What is typical for call centre jobs? job characteristics, and service interactions in different call centres. *European journal of work and organizational psychology* **12**, 311–340, https://doi.org/10.1080/13594320344000183 (2003).
37. Jeong, K.-S., Choi, S.-J., Park, M.-O. & Li, Y. The effects of customer service representatives' emotional labor by emotional display rules on emotional dissonance, emotional exhaustion and turnover intention in the context of call centers. *Korean J Bus Adm* **28**, 529–51 (2015).
38. Gao, J., Leung, H. K., Fan, J., Wu, B. W. Y. & Sik, H. H. The neurophysiology of the intervention strategies of awareness training program on emotion regulation. *Frontiers in Psychology* **13**, 891656, https://doi.org/10.3389/fpsyg.2022.891656 (2022).
39. Schneider, J. *et al*. The role of face masks in the recognition of emotions by preschool children. *JAMA pediatrics* **176**, 96–98, https://doi.org/10.1001/jamapediatrics.2021.4556 (2022).
40. Kroenke, K., Spitzer, R. L. & Williams, J. B. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine* **16**, 606–613, https://doi.org/10.1046/j.1525-1497.2001.016009606.x (2001).
41. Cohen, S., Kamarck, T. & Mermelstein, R. A global measure of perceived stress. *Journal of health and social behavior* 385–396, https://doi.org/10.2307/2136404 (1983).
42. Luthans, F., Avolio, B. J., Avey, J. B. & Norman, S. M. Positive psychological capital: Measurement and relationship with performance and satisfaction. *Personnel psychology* **60**, 541–572, https://doi.org/10.1111/j.1744-6570.2007.00083.x (2007).
43. John, O. P., Srivastava, S. *et al*. The big-five trait taxonomy: History, measurement, and theoretical perspectives. (1999).
44. Rosenberg, M. *et al*. *Society and the adolescent self-image*, vol.11 (Princeton university press Princeton, NJ, 1965).
45. Diener, E., Emmons, R. A., Larsen, R. J. & Griffin, S. The satisfaction with life scale. *Journal of personality assessment* **49**, 71–75, https://doi.org/10.1207/s15327752jpa4901_13 (1985).
46. Kim, J.-H., Kim, B.-H. & Ha, M.-S. Validation of a korean version of the big five inventory. *Journal of Human Understanding and Counseling* **32**, 47–65 (2011).
47. McColl-Kennedy, J. R., Patterson, P. G., Smith, A. K. & Brady, M. K. Customer rage episodes: emotions, expressions and behaviors. *Journal of Retailing* **85**, 222–237, https://doi.org/10.1016/j.jretai.2009.04.002 (2009).
48. Surachartkumtonkun, J., McColl-Kennedy, J. R. & Patterson, P. G. Unpacking customer rage elicitation: A dynamic model. *Journal of Service Research* **18**, 177–192, https://doi.org/10.1177/1094670514556275 (2015).
49. Harris, L. C. Service employees and customer phone rage: An empirical analysis. *European Journal of Marketing* **47**, 463–484, https://doi.org/10.1108/03090561311297418 (2013).
50. Schuller, B. & Batliner, A.*Computational paralinguistics: emotion, affect and personality in speech and language processing* (John Wiley & Sons, 2013).
51. Akçay, M. B. & Oğuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication* **116**, 56–76, https://doi.org/10.1016/j.specom.2019.12.001 (2020).
52. Koduru, A., Valiveti, H. B. & Budati, A. K. Feature extraction algorithms to improve the speech emotion recognition rate. *International Journal of Speech Technology* **23**, 45–55, https://doi.org/10.1007/s10772-020-09672-4 (2020).
53. Ardila, R. *et al*. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4218–4222 (2020).
54. Burkhardt, F. *et al*. A database of german emotional speech. In *Interspeech*, vol.5, 1517–1520 (2005).
55. Giannakakis, G. *et al*. Stress and anxiety detection using facial cues from videos. *Biomedical Signal Processing and Control* **31**, 89–101, https://doi.org/10.1016/j.bspc.2016.06.020 (2017).
56. Giannakakis, G., Manousos, D., Chaniotakis, V. & Tsiknakis, M. Evaluation of head pose features for stress detection and classification. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 406–409, https://doi.org/10.1109/BHI.2018.8333454 (2018).
57. Lee, D., Park, E., G., K., Han, Y. & Lee, U. Emowork: A multimodal dataset for assessing emotion, stress, and emotional workload in interpersonal work scenario, https://doi.org/10.5281/zenodo.15181219 (2025).
58. Sharma, K., Castellini, C., Van Den Broek, E. L., Albu-Schaeffer, A. & Schwenker, F. A dataset of continuous affect annotations and physiological signals for emotion analysis. *Scientific data* **6**, 196, https://doi.org/10.6084/m9.figshare.9891446 (2019).
59. Shojaei, N. *et al*. A publicly available pharyngitis dataset and baseline evaluations for bacterial or nonbacterial classification. *Scientific Data* **12**, 1418, https://doi.org/10.1038/s41597-025-05780-5 (2025).
60. Breiman, L. Random forests. *Machine learning* **45**, 5–32, https://doi.org/10.1023/A:1010933404324 (2001).
61. Chen, T. *et al*. Xgboost: extreme gradient boosting. *R package version 0.4-2* **1**, 1–4 (2015).
62. Loh, W.-Y. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery* **1**, 14–23, https://doi.org/10.1002/widm.8 (2011).
63. Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. & Scholkopf, B. Support vector machines. *IEEE Intelligent Systems and their applications* **13**, 18–28, https://doi.org/10.1109/5254.708428 (1998).

64. Duda, R. O., Hart, P. E. *et al*. *Pattern classification and scene analysis*, vol.3 (Wiley New York, 1973).
65. Park, E., Lee, D., Han, Y., Diefendorff, J. & Lee, U. Hide-and-seek: Detecting workers' emotional workload in emotional labor contexts using multimodal sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.***8**, https://doi.org/10.1145/3678593 (2024).
66. Welch, G., Bishop, G. *et al*. An introduction to the kalman filter. (1995).
67. Wusk, G. C., Abercromby, A. F. & Gabler, H. C. Psychophysiological monitoring of aerospace crew state. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, 404–407, https://doi.org/10.1145/3341162.3349309 (2019).
68. Linssen, L., Landman, A., van Baardewijk, J. U., Bottenheft, C. & Binsch, O. Using accelerometry and heart rate data for real-time monitoring of soldiers' stress in a dynamic military virtual reality scenario. *Multimedia Tools and Applications* **81**, 24739–24756, https://doi.org/10.1007/s11042-022-12705-6 (2022).
69. Han, Y., Toshnazarov, K. E., Kim, B. H., Noh, Y. & Lee, U. Watchppg: An open-source toolkit for ppg-based stress detection using off-the-shelf smartwatches. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, UbiComp/ISWC '23 Adjunct, 208-209, https://doi.org/10.1145/3594739.3610800 (Association for Computing Machinery, New York, NY, USA, 2023).
70. Elfenbein, H. A. & Ambady, N. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin* **128**, 203, https://doi.org/10.1037/0033-2909.128.2.203 (2002).
71. Kim, Z. H. & Lausberg, H. Koreans and germans: Cultural differences in hand movement behaviour and gestural repertoire. *Journal of Intercultural Communication Research* **47**, 439–453, https://doi.org/10.1080/17475759.2018.1475296 (2018).
72. Laukka, P. & Elfenbein, H. A. Cross-cultural emotion recognition and in-group advantage in vocal expression: A meta-analysis. *Emotion Review* **13**, 3–11, https://doi.org/10.1177/1754073919897295 (2021).
73. Etzion, D. & Pines, A. Sex and culture in burnout and coping among human service professionals: A social psychological perspective. *Journal of cross-cultural Psychology* **17**, 191–209, https://doi.org/10.1177/0022002186017002004 (1986).
74. Greco, A., Valenza, G., Lanata, A., Scilingo, E. P. & Citi, L. cvxeda: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering* **63**, 797–804, https://doi.org/10.1109/TBME.2015.2474131 (2016).
75. Russell, J. A. A circumplex model of affect. *Journal of personality and social psychology* **39**, 1161, https://doi.org/10.1037/h0077714 (1980).
76. Park, S.-J., Choi, H.-R., Choi, J.-H., Kim, K.-W. & Hong, J.-P. Reliability and validity of the korean version of the patient health questionnaire-9 (phq-9). *Anxiety and mood* **6**, 119–124 (2010).
77. Lee, J. *et al*. The reliability and validity studies of the korean version of the perceived stress scale. *Korean Journal of Psychosomatic Medicine* 127–134 (2012).
78. Lim, T.-H. Validation of the korean version of positive psychological capital (k-ppc). *Journal of coaching development* **16**, 157–166 (2014).
79. Lee, J., Nam, S., Lee, M. k., Lee, J.-h. & Lee, S.M. Rosenberg'self-esteem scale: Analysis of item-level validity. *The Korean Journal of Counseling and Psychotherapy* **21**, 173–189 (2009).
80. Lim, Y. J. Psychometric properties of the satisfaction with life scale among korean police officers, university students, and adolescents. *Korean Journal of Psychology: General* **31**, 877–896 (2012).

Acknowledgements

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2022-0-00064, Development of Human Digital Twin Technologies for Prediction and Management of Emotion Workers' Mental Health Risks) and the National Research Foundation of Korea (NRF) funded by the MSIT (RS-2022-NR068758).

Author contributions

Duri Lee designed, prepared, and conducted the data collection; constructed and pre-processed the dataset; performed the technical validation; and wrote the manuscript and all associated results. Eunji Park co-designed, prepared, and conducted the data collection; contributed to dataset construction and pre-processing; participated in code development; and revised the manuscript. Gyuna Kim contributed to dataset construction and revised both the manuscript and the code. Yunjo Han assisted in data collection, contributed to pre-processing. Uichin Lee supervised the dataset design and data collection, provided overall project guidance, and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to U.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026