Hide-and-seek: Detecting Workers' Emotional Workload in Emotional Labor Contexts Using Multimodal Sensing

EUNJI PARK, Chung-Ang University, Republic of Korea
DURI LEE, KAIST, Republic of Korea
YUNJO HAN, KAIST, Republic of Korea
JAMES DIEFENDORFF, University of Akron, United States
UICHIN LEE*, KAIST, Republic of Korea

Emotional labor refers to the process in which workers are required to express emotions regardless of their actual feelings by the organization. In workplaces where such display rules exist, workers experience an emotional workload. Continued exposure to emotional workload can lead to severe mental and psychological issues. Nevertheless, research on assessing emotional workload remains understudied. In this study, we propose a machine learning model to automatically evaluate workers' emotional workload in emotional labor situations through multimodal sensing. The data collection study was designed based on a call center scenario. Within the study, we manipulated customer behaviors as confederates and assessed the worker's emotional workload. As a result, this study provides a benchmark using well-known features and standard machine learning methods. We achieved an accuracy of up to 87% for binary and three-class classification cases. Finally, we discuss the significance of assessing emotional workload and considerations for its practical application in the workplace.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **Laboratory experiments**.

Additional Key Words and Phrases: Emotion Regulation, Emotional Workload, Multimodal Sensing

ACM Reference Format:
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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 119 (September 2024), 28 pages. https://doi.org/10.1145/3678593

1 INTRODUCTION

Customer-facing workers are often required to express positivity to customers, regardless of their actual emotions, which is known as a 'display rule' in emotional labor. In accordance with these display rules, the workers need to engage in constant self-control to regulate their own emotions. For example, in a call center, workers are expected to remain courteous even when dealing with customers who may shout and use offensive language, requiring them to suppress negative emotions. However, since self-control capacity is a limited resource [9], continuously regulating their own emotions can lead to serious psychological and mental issues for workers. This aspect of constantly regulating their own emotions as part of their job requires considerable mental effort, which we referred to as 'emotional workload' in this study.

*Corresponding author

Authors' addresses: Eunji Park, Chung-Ang University, Seoul, Republic of Korea; Duri Lee, KAIST, Daejeon, Republic of Korea; Yunjo Han, KAIST, Daejeon, Republic of Korea; James Diefendorff, University of Akron, Rocquencourt, United States; Uichin Lee, KAIST, Daejeon, Republic of Korea.

2474-9567/2024/9-ART119
https://doi.org/10.1145/3678593

Workload, typically associated with performing regular tasks, has been actively studied in the field of Human-Computer Interaction (HCI). However, until now, it has mainly been limited to measuring cognitive workload in situations where multitasking is required, such as workplaces that involve knowledge-based tasks [94] or machine maneuvering tasks (e.g., driving and flight control) [30, 113]. Emotional workload differs from cognitive workload in that it has an interpersonal nature (e.g., emotional workload relies on stimuli from social interactions such as conversations with customers). Therefore, the person with whom one is interacting is also considered an important factor for measuring emotional workload. Topics related to emotional workload have been actively discussed in the field of psychology. Psychology researchers developed various underlying theoretical models of the emotional regulation process, including stimuli from interaction partners (e.g., customer outbursts), emotions demanded by the workplace (i.e., display rules), and the emotions actually experienced by the worker. The mental effort for emotional work that occurs during this process has been measured by questionnaires [39, 86]. The theoretical models provide a foundation for understanding and evaluating emotional workload. However, existing psychological studies have relied on self-report, and studies that automatically assess emotional workload are still lacking. Motivated by this, our study focuses on emotional workers in customer contact centers and addresses the following research questions:

- RQ1: How can we model the emotional workload of workers in workplaces that require emotional labor?
- RQ2: What are the dominant features for automatically recognizing emotional workload by using multimodal sensing?

In order to address the research questions, we built machine learning models to automatically assess workers' emotional workload in real-time within the context of emotional labor through multimodal sensing. To answer the first research question, we defined *emotional workload* based on existing theoretical models and verified the performance of the machine learning models built using various classifiers. We designed a data collection study and employed a call center scenario, a widely used protocol for observing emotion regulation [31]. We collected multimodal sensor data that includes stimuli and worker's response under a display rule. To implement stimulus, we manipulated the customer's behavior based on the scenario and observed changes in the worker's emotional workload. The customer made a total of three calls per worker, and the customer was asked to act as a neutral, shouting, or swearing customer in each randomly given call condition. Simultaneously, we measured workers' responses through wearable sensors that can capture physiological response such as heart and electrodermal activities. Furthermore, we collected voice data from both the customer and the worker during their conversations. Each call condition lasted for approximately 4 minutes, and we collected a total of 12 minutes of data per worker.

To construct machine learning models, we employed *stimulus-based labeling* method commonly utilized in previous affective computing research [95]. We binary labeled based on the intensity of the given stimulus under conditions of low- and high-stimulus levels. Furthermore, we expanded the labeling to three classes by categorizing stimulus types (e.g., shouting customer, swearing customer) along with a neutral condition for comparison. We also verified the performance of the models utilizing binary and three-class labels derived from workers' self-reported emotional workload. Machine learning models were created using a total of seven classifiers, and validation was conducted through Leave-One-Subject-Out (LOSO) cross-validation. To answer the second research question, we analyzed the importance of different types of features. Features were categorized into situational cues (customer's voice), required emotional responses (worker's voice), and not-required emotional responses (worker's physiological data), following a theoretical model of emotion regulation. We compared the model's performance using various combinations of these features. In addition, we conducted feature importance analysis using SHapley Additive ex-Planation (SHAP) [69] .

As a result, we revealed that modeling based on stimulus-based given emotional workload performs better than modeling based on self-reported perceived emotional workload. For binary classification, the model using a Support Vector Machine (SVM) classifier demonstrated the highest performance with an accuracy of 0.87. In the

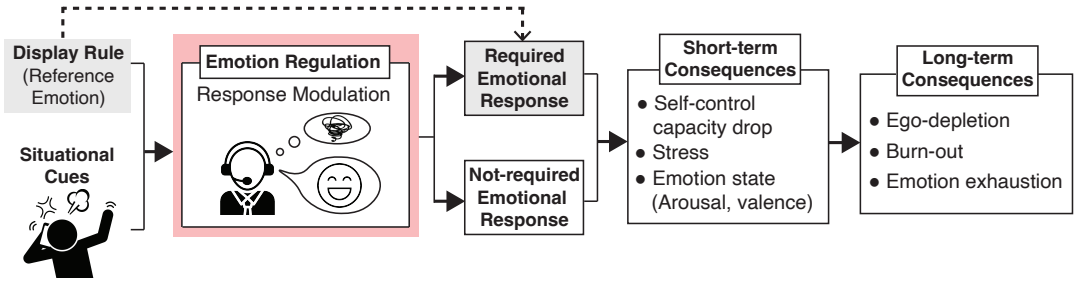

Fig. 1. Theoretical Model of Emotion Regulation: Under the display rule, workers regulate their actual felt emotion and express the required emotion in the emotional labor context. In the short term, it reduces self-capacity, induces stress, and alters emotional states, while in the long term, it leads to serious psychological and mental issues such as ego-depletion, burnout, and emotional exhaustion.

case of three-class classification, the accuracy of the model based on given emotional workload and using an SVM classifier reached 0.82. Furthermore, we observed that the model's performance was higher when utilizing not-required emotional response (worker's physiological data) and situational cue (customer's voice) features compared to required emotional response (worker's voice) features. The model that solely utilized situational cue features achieved the highest accuracy at 0.88. In contrast, the accuracy of the model using only required emotional response features was 0.65. Adding required emotional response features to other feature combinations did not lead to performance improvement. Based on the results, we propose a model that accurately measures the emotional workload of workers in workplaces where emotional labor is required. Furthermore, we discuss important considerations when applying this research in actual workplace settings.

2 BACKGROUND AND RELATED WORK

2.1 Theoretical Backgrounds on Emotion Regulation and Workload

2.1.1 Theoretical Backgrounds on Emotion Regulation. Emotion regulation refers to the process of individuals adjusting their emotions in order to function more effectively in the workplace or behave appropriately in social situations. In the field of psychology, numerous studies have been conducted on understanding and modeling emotion regulation. Previous psychological studies explored how individuals regulate their emotions in terms of mechanisms, reasons, and consequences [20, 37, 41, 42, 110]. In particular, emotion regulation has been regarded as an important research topic in service-related workplaces involving emotional labor [66, 104]. In customer service centers, employees are required to follow *display rules* [35], whereby expressing positive emotion to satisfy the customers. In other words, emotional workers are expected to suppress their own negative emotions and respond courteously, even when customers yell, swear, or express strong complaints. In this regard, Grendey proposed a framework for emotion regulation in workplace contexts that impose specific emotions through such *display rules* [37]. According to this framework, emotion regulation in the workplace is influenced by situational cues such as emotional events, as illustrated in Figure 1. After perceiving the situational cue, employees regulate their emotions according to display rules, which is known as *emotion acting*. As such, managing and modifying emotions in the workplace has been considered a part of the job and responsibilities such as in customer call centers [37], and consistently regulating their emotions can itself be perceived as a workload by the workers. There is a growing interest in exploring emotion regulation in the HCI and Ubicomp community; e.g., understanding the use of emotion AI in emotional work contexts [92] and the influence of digital technologies for everyday emotion regulation [99, 108].

2.1.2 Theoretical Backgrounds on Workload. While there is no universally agreed-upon definition for workload, it has been employed to explain the inability of a human operator to cope with the requirements of a task in situations where human capacity is finite [36]. Kramer also defined the workload as the cost of performing a task in terms of a reduction in the capacity to perform additional tasks that use the same processing resource [63]. As such, the various definitions of workload commonly emphasize that workload depletes the finite capacity resource due to the mental effort demands of the task. Depending on the work domain and the nature of tasks to be performed, workplaces encompass various types of workloads. For knowledge workers, one of the typical workloads caused by the number of tasks and time constraints is cognitive load. Cognitive load refers to the amount of information that working memory can process for a given time [102], and analyzing cognitive load is particularly useful in knowledge work contexts. For example, research on estimating the cognitive load of pilots or drivers, who need to perform multiple tasks simultaneously within limited time, has been actively conducted for safety purposes [5, 30]. Another aspect of the workload is *emotional workload* which can arise from subjective psychological experiences [70], particularly when users engage in emotional labor. Emotional workload is proportional to the extent of required emotion regulation [39, 117]; e.g., high levels of arousal are more difficult to regulate. So far, there have been many studies to measure cognitive load at the workplace, but prior studies on measuring emotional workload are lacking.

Traditionally, the measurement of cognitive workload has been conducted by measuring a person's mental effort while performing a given task [84]. In general, the measurement of mental effort is composed of three methods: (1) subjective ([18, 47, 90], (2) psychophysiological, and (3) task- and performance-based indices [28]. Subjective techniques assume that humans can introspect their cognitive processes and report the amount of mental effort they have expended. Psychophysiological techniques obtain information about mental effort based on physiological indices such as pupillary diameter and heart-rate variability. Task- and performance-based techniques directly measure the task's performance itself, such as acquisition time and number of errors. Emotional workload shares a commonality with cognitive load in that both consume mental effort to engage in emotion regulation or information processing. *Mental effort for emotional work* is different from that for cognitive work, and thus, dedicated measures based on self-reports [116] were proposed in the literature, such as Emotion Regulation Questionnaire and Difficulties in Emotion Regulation Scale [6, 33, 34, 39, 43, 56, 83, 86, 88, 103]. Furthermore, researchers attempted to assess *fine-grained mental efforts* for emotion regulation while a worker is performing a task (e.g., efforts for suppressing felt emotions), by employing a monetary assessment approach [31] where a worker reviews a recorded video and retrospectively annotate mental efforts (say in every 200 ms).

Similar to how cognitive load depletes human attention resources and leads to reduced task performance [111, 112], it is important to note that emotional workload depletes a worker's self-control capacity and eventually exhausts an emotion reservoir (known as ego depletion [9, 10]). This means that continuous self-tracking of emotional workload will offer novel opportunities for workers to better manage their emotional health and well-being in their workplaces (e.g., taking some rest when the cumulative workload is too high). However, self-tracking emotional workload via self-reports is challenging because workers must provide their responses after completing tasks, or reviewing a recorded video for retrospective annotation, which incurs significant burden. In this work, we aim to alleviate the burden on workers regarding self-report measures, by automatically assessing the mental effort required for emotion regulation in real time.

2.2 Automatic workload and Affect assessment in HCI

In the field of HCI, much research aims to automatically estimate cognitive workloads and emotional states using sensor data. Cognitive workload assessment studies have been conducted in situations requiring high cognitive workload, such as flying, driving, knowledge work, manufacturing, and emergency management

contexts [30, 60, 77, 94, 113, 119]. Affective computing studies involve inferring emotions or stress states by analyzing users' responses to emotionally evocative stimuli using music, videos, or images [13, 44, 82, 100, 109].

Cognitive workload assessment leverages sensor data on a user's behavioral and physiological responses collected while performing real-world tasks (e.g., road driving [30], flight [113]) or basic cognitive tasks that represent cognitive aspects of real tasks (e.g., math calculations, recall tasks). These tasks are then used as 'given ground truths' for modeling. Note that users can also self-report perceived cognitive workload while performing such tasks, by answering questionnaires such as NASA-TLX [45], or a single item of task difficulty [17], and thus, perceived cognitive workload can alternatively be used as ground truth for cognitive workload modeling. For workload sensing, so far researchers have leveraged various behavioral response data, such as reaction time [30] and speech features [60, 114] and physiological response data, such as heart rate (HR), electroencephalogram (EEG), and pupil response [17, 30, 45, 77, 113]. Such sensor data are then used to build machine learning models that can automatically measure cognitive workload status, ranging from simple linear models [72] to complex neural networks [17].

Similarly, prior studies on automated stress and emotion detection leveraged collected data on behavior and physiological responses while users were performing stressful or emotionally stimulating tasks [95]. Task types (i.e., stimulus and no-stimulus conditions) were then used as ground truths to build a machine learning model. After collecting response data to stimuli, researchers often asked the participants to self-report their perceived emotional states, for example, by answering a single item Likert scale question for stress level [49, 52, 105], or valence-arousal questionnaires based on Russell's circumplex model [62, 78, 82, 109]. As in workload assessment, prior studies on stress and emotion detection collected physiological data, including an electrocardiogram (ECG), Electrodermal activity (EDA), and EEG [78, 82, 85, 95, 105, 109], and behavioral data, including video data such as facial expression and posture, as well as audio features [2, 61, 78, 97]. As for behavioral response data, several studies analyzed voice data, extracting various prosodic features such as fundamental frequency (F0), energy, voiced_flag, Mel-frequency cepstral coefficients (MFCCs), spectral features, and Zero Crossing Rate (ZCR) for emotion classification [2, 61, 97]. Using features extracted from behavioral and physiological response data, regression and classification models that recognize and predict stress and emotion are built [95, 105, 109]. As such, various methods for automatically evaluating cognitive workload or emotional state have been dealt with in existing HCI literature, but to our knowledge, there is no prior work that measured emotional workload reflecting mental effort for emotion regulation.

Thus, the goal of this work is to automatically classify emotional workload for emotional labor. While existing research on stress assessment relies on social stress tests, which involve memorized, face-to-face presentations, our approach adopts a different scenario for emotional workload measurement. Specifically, we simulate realistic emotional work situations of emotional labor such as a customer service call scenario as studied by Gabriel and Diefendorff [31]. Based on the findings about the stressful conditions that call center workers encounter during the call with the customers [75], we designed realistic scenarios that require varying levels of emotional effort to comply with display rules. Having *multiple levels of emotional workload* is a major departure from existing 'stress' studies in which baseline conditions assume a lack of stimulus (i.e., resting states). Our research design addresses this gap by simulating realistic emotional workload when the same display rules are required in all conditions. As in prior studies on automatic cognitive and affective assessment, we demonstrate the feasibility of automatically measuring emotional workload, by leveraging physiological and behavioral data collected from realistic emotional work scenarios.

3 DATA ACQUISITION

For realistic emotional workload assessment, we considered a data collection scenario, following prior studies that involve interpersonal interaction with *high-impact manipulations (i.e., confederate)* of human behavior to

create realistic emotion-eliciting situations [4]. We adopted the conventional methods for emotion elicitation that involves interpersonal interaction between customers and employees in a call center situation [31, 46, 75, 101]. We used a within-episode data collection scenario similar to real-life interactions in a call center. Also, this study has received ethical approval from the university Institutional Review Board (IRB).

3.1 Participants

We recruited a total of 31 participants (female: 24, male: 7) who are currently working as emotion workers at six companies operating call centers. Since the gender composition of the workforce in this profession is largely skewed to female [12, 16] (e.g., over 90% in Korea [55]), 77.4% of participants were female. The average work experience of the participants was 3.25 years (SD= 2.11). On the day of the data collection study, we requested that participants refrain from engaging in activities that could affect their physiological responses, such as consuming alcohol, smoking, engaging in intense exercise, or consuming caffeine [95].

We opted to recruit a professional actors for the role of the customer from the Local Theater Association, to ensure the expression of specific emotions within a concise time frame, and to maintain consistency throughout the data collection study, as in prior emotion regulation studies [31, 35]. In particular, our experiment was based on realistic emotion-eliciting situations, so recruiting professional actors was crucial to provide workers with an immersive environment through the actor's natural acting in interpersonal interaction situations. To avoid a gender effect, we recruited 2 male actors and 2 female actors. One customer acted in all sessions, thereby guaranteeing a consistent voice tone, conversation content, and emotional expression throughout the whole duration of our study.

3.2 Tasks

Figure 4 illustrates how we organize call sessions and how we define labels for emotional workload assessment. Defining labels is critical for emotional workload assessment. In this work, we consider two different approaches for workload labeling: i.e., (1) 'given emotional workload' conditions as labels (e.g., neutral, shouting, or swearing conditions), and (2) a worker's 'perceived emotional workload' based on self-reports as labels (e.g., low and high workload).

3.2.1 Call session. In this data collection, emotion workers and customer engaged in three telephone conversations, each lasting about four minutes, to resolve a given issue. During the three calls, the workers engaged in continuous conversations with the same customer in the same problem situation. The customer acted with different attitudes based on the given scenario for each call to express dissatisfaction, including 'Customer with neutral voice tone', 'Shouting customer', and 'Swearing customer'. Shouting and swearing are the typical customers' rage-associated behavior [75], and aggressive verbal expression such as shouting and swearing that violates social norms influences the emotional workload of workers [38]. In real-life situations, customers may mix aggressive speech expressions. However, there are cases where certain speech expressions are prominent. Therefore, in this study, we selected the specific behaviors of swearing and shouting and distinguished between situations where customers' anger expression behaviors are prominent in terms of vocal features such as pitch in shouting, and situations where customers' anger is conveyed through expression in swearing, although pitch features are not prominent.

In the neutral voice tone scenario, the customer asks the representative questions in a neutral tone, seeking factual information about the situation. In the shouting session, we demanded that the customer raise their voice and shout at the representative without using anger-inducing words or profanity. In the swearing session, the customer was instructed to convey their demands to the worker using anger-inducing words known to evoke frustration. We provided the customer with examples of words that could be used during the call. However, to clearly differentiate between shouting sessions and swearing sessions, we requested customers not to shout

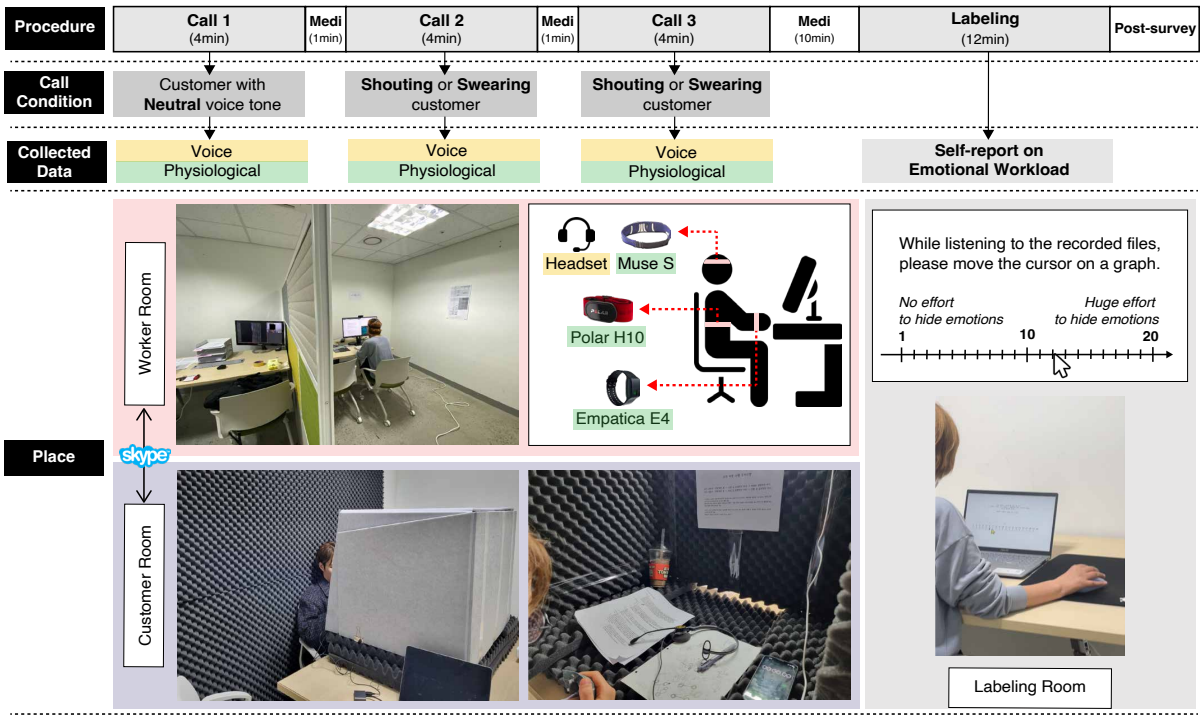

Fig. 2. Data Acquisition (Comprehensive Process of Call Sessions and Labeling Session for Emotional Workload Assessment): Call sessions involve customer interactions with workers in various scenarios, where customer attitudes range from neutral to shouting and swearing. Workers' voice and physiological data were collected throughout these interactions. The labeling session illustrates the retrospective affect annotation protocol used to collect self-report labels, tracking the mental effort for emotion regulation of participants within a short time frame.

in swearing sessions. In addition, any excessive profanity or expressions of sexual harassment were strictly prohibited throughout all conditions. To simulate the realistic work environment where customer information is accessible, we provided specific numerical data to the customer such as the customer's card usage amount, item purchase history, and the company's specified cashback requirements. We also instructed the workers to use examples of opening and closing statements and follow a *display rule* [35], where they were asked to maintain kindness as much as possible towards the customer regardless of the situation.

Following prior studies [31, 35], we carefully developed scenario-based naturalistic tasks based on common situations that occur in customer service centers. After consulting with local call centers, the following scenario was designed: A credit card company organized a promotional event to provide cashback to customers who recently opened a new card. A customer failed to receive a reward due to a requirement misunderstanding. The customer called the card company's customer service center to complain about failing to receive the cashback.

3.2.2 Labeling session. The labeling method we employed is a *retrospective affect annotation protocol* which has been widely used in the affective computing field for self-annotation [1, 27]. After the call session ended, we asked workers to listen to recorded audio files of the conversations and label the mental effort exerted for emotion regulation at that time to measure the emotional workload of workers during the calls. According to studies explaining the emotion process, experienced and expressed emotions can vary significantly and show

substantial variability within a short time frame [65, 73, 74]. In addition, the theoretical explanations of emotional labor noted the importance of within-episode dynamics in affective computing [11, 19, 22, 23, 89]. Therefore, to accurately capture the changing states of the participants within a short period, we collected momentary and continuous labeling data. We measure emotional workload by asking how much participants made an effort to hide true emotions they felt during the call [31]. The E-prime 3.0 Professional software that has been commonly used for conducting psychological experiments [32] was used for retrospective affect judgment [26]. As shown in Figure 4, the participants responded by moving the mouse cursor left or right on a graph in the form of a 1D bar, indicating their state during the call (1: 'no effort made to hide my true emotions', 20: 'a great deal of effort made to hide my true emotions'). We used a scale of self-report labeling from 1 to 20, by referring to prior studies that attempted a momentary approach to measure emotion regulation [31].

3.2.3 Apparatus. We collected sensor data from four devices. Firstly, we used a chest-worn device, Polar H10, to monitor heart activities. The workers placed the sensor directly above their skin near the chest area. Using the Polar H10 sensor, we measured an electrocardiogram (ECG) at 130Hz. The data, including ECG raw data, HR, and RR intervals, were saved to a designated local storage via the Polar Beat application using Bluetooth. Secondly, we used the wrist-worn device, Empatica E4 band, to collect data such as blood volume pulse (BVP, 64Hz), electrodermal activity (EDA, 4Hz), temperature (TEMP, 4Hz), acceleration (ACC, 32Hz), and interbeat interval (IBI). The workers were instructed to wear the Empatica E4 band on their non-dominant hand, ensuring the sensor remained in contact with the skin tight enough without causing discomfort. To synchronize the data with the other devices, a mobile application was developed based on empalink-sample-project-android code to log timestamps at each data collection interval. The Empatica E4 band data was transmitted to a mobile device via Bluetooth for logging. Thirdly, we used Muse S to collect the workers' electroencephalogram (EEG) signals. Muse S is a headband device with four electrodes that collect EEG signals from TP9, AF7, AF8, and TP10 positions at a rate of 256Hz. The EEG data was sent from the Muse S device to the local computer via Bluetooth. An EEG data logging system was implemented using the muse-lsl Python package, which allows for streaming, visualizing, and recording Muse S data. Lastly, the workers wore a headset microphone while conducting tasks, and we recorded their voices (44.1kHz). To provide an environment similar to their usual workplace counseling situations, we used the same headset model that the workers actually used in one of their three workplace locations. Additionally, we also collected the voices of the customer. To ensure similar audio quality, the customer wore the same headset as the worker-participants during the data collection. As a result, the emotion workers wore four devices (Polar H10, Empatica E4 band, Muse S, and Headset). Two Galaxy Note 9 devices were used for data logging; i.e., one device collected data from the Polar H10 via the Polar Beat application, and another device collected data from the Empatica E4 band. The EEG data and audio data were collected and stored on a local device wirelessly connected to the computer.

3.2.4 Setup and Procedure. Figure 4 demonstrates the setup and procedure of data collection. Prior to the start of the data collection, preparations were made in both the worker and customer rooms. In the worker room, a computer for conducting the calls was set up on a regular desk, and there was a separate space next to the worker for researchers to monitor the data logging process. Upon arrival at the lab, each worker was instructed to sit in front of the computer, and an explanation of the data collection procedure and scenario was provided. The customer room also had a computer on a regular desk used for making calls. Before each task, instructions for the role-playing and specific requirements for each call condition were provided to ensure consistent content and voice from the customer. Printed instructions were also posted inside the soundproofing materials for the customer to refer to during the data collection. The customer wore the same headset as the workers and waited until the other party is ready. Once the preparation was finished, the customer initiated the calls to the worker via Skype. To achieve a naturalistic scenario and conversational flow, the first session was fixed as a neutral voice tone session, while the shouting and swearing sessions were set randomly. After customer ended the call with the

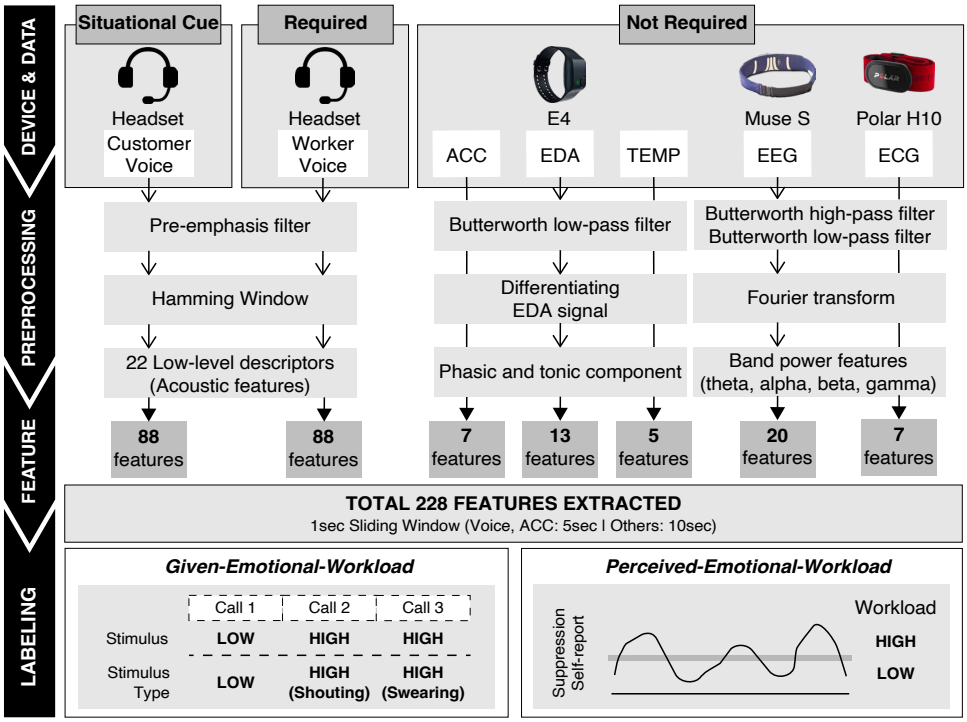

Fig. 3. Data Analysis Methodology. The process initiates with the comprehensive collection and synchronization of various data types, encompassing audio, physiological (EDA, EEG, TEMP, and ECG), and acceleration data. Subsequently, preprocessing procedures employ specialized techniques for each data type, resulting in the extraction of a rich array of features for the analysis. Further, we describe two distinct data labeling methodologies: the 'Given-Emotional-Workload' approach, which assesses workload based on predefined stimuli types and intensities, and the 'Perceived-Emotional-Workload' approach, wherein worker self-reported labels are categorized using a threshold set as the average of values provided by each worker.

worker, the worker performed 1 minute of meditation. During the meditation, the worker wore the headset and watched a mindful breathing with relaxing background music. After the completion of the last call, the worker took off the worn devices and engaged in 10 minutes of meditation to alleviate any psychological stimulation received during the task. After the meditation, the worker moved to the labeling room. The labeling process was performed with a computer while sitting at a desk. We provided the participants with self-report instructions. Participants conducted *practice session* asking them to use the E-prime interface three times for being familiar with the tools. Once the practice session was completed, the worker listened to the conversation recordings and performed ratings for suppress. In the post-survey after completing labeling, we evaluated the validity of our scenario by asking participants about the perceived stress of workers from each task, how realistic the scenarios were, and how stressful the lab environments were to close the data collection. Participants answered significantly higher levels of stress on call 2 and call 3 than on call 1 and, 29 out of 31 participants marked that the given call situations and clients were very realistic. Regarding the lab environments, participants answered that they got stress about the situation to collect data and wearing the equipment, but at a very low level.

Table 1. List of extracted features

Data	Feature Type	# of Features	Description
Audio	F0	4	
	Energy	4	
	ZCR	4	
	Voiced-flag	4	
	13 MFCCs	52	
	Spectral centroid	4	mean, standard deviation, min, max values
	Spectral bandwidth	4	
	Roll-off	4	
	Roll-off min	4	
	Spectral contrast	4	
EDA	Raw signal	4	mean, standard deviation, min, max values
	Phasic component	4	mean, standard deviation, min, max values
	Tonic component	4	mean, standard deviation, min, max values
	Peak	1	number of peaks detected
EEG	Theta power	4	
	Alpha low power	4	
	Alpha high power	4	mean, standard deviation, min, max values
	Beta high power	4	
	Gamma high power	4	
TEMP	Body temperature	5	mean, standard deviation, min, max, slope values
ECG	HRV	5	meanRR, BPM, SDNN, RMSSD, PNN50
	HR	2	mean, standard deviation values
ACC	Raw signal of each axis (x,y,z)	6	mean, standard deviation values
	Magnitude from 3 axes	1	mean magnitude value

4 DATA ANALYSIS METHODOLOGY

Figure 3 presents the data analysis methodology for the modeling in our study.

4.1 Pre-processing and Feature Extraction

We collected audio data of the voices of the customer and emotional workers, physiological data (EDA, EEG, TEMP, and ECG), and acceleration data. To synchronize the various types of data, we collected epoch times in milliseconds as timestamps for each data type and merged the data based on the timestamps. Previous studies on emotion detection typically used a 60-second window [64], as they provided a similar level of stimulation to participants within a single condition. In our study, we considered the dynamically changing contexts of the ongoing conversation with different stimulations given to participants (e.g., shouting or swearing), and thus, we selected a 10-second window except for audio and acceleration data. In actuality, we observed that the model's performance slightly improved when the window for physiological data was set to 10 seconds compared to setting it to 60 seconds (See Section 5.1.1). The 5-second window was chosen for audio and acceleration data based on the previous research [95], and we averaged the label data with a 5-second window as well. Using a sliding window of 1 second, we extracted features from the data, and they are displayed in Table 1. All features were normalized using z-normalization for each subject before training the models [81].

4.1.1 Audio data. For audio data, we first applied a standard pre-emphasis filter for energy normalization without downsampling to extract low level descriptors (LLD). As typical prosodic features can capture subtle changes in vocal features [97], we employed a frame size of 25m/s with 10m/s overlapping (Hamming window function). We extracted 22 acoustic features typically used in speech analysis, including the fundamental frequency (F0), energy, zero crossing rate (ZCR), voiced_flag, 13 Mel-frequency cepstral coefficients (MFCCs), and 5 additional spectral features (centroid, bandwidth, rolloff, rolloff_min, and contrast) by using the librosa library in Python.

Then, we extracted the mean, standard deviation, min and max values from each LLD per 5-second window, resulting in a total of 88 features.

4.1.2 Physiological data. EDA data: To remove high-frequency noise from the EDA signal, we used a first-order Butterworth low-pass filter with a cutoff frequency of 0.4 Hz. We extracted the mean, standard deviation, min and max values of the signal every 10 seconds as a total of 4 features. Additionally, we decomposed the EDA signal into its tonic component, representing the slowly varying baseline conductivity, and the phasic component, representing the short-term response to stimuli [95]. For the decomposition, we utilized a convex optimization approach proposed by Greco et al [40]. After decomposition, we extracted the mean, standard deviation, min and max values of phasic and tonic signals every 10 seconds as a total of 4 features. Furthermore, we utilized the number of peaks in the phasic component as a feature. Counting the number of phasic component peaks over a specific time period is a widely used method for measuring sympathetic arousal [21]. To remove noise from the phasic signal, we applied a Bartlett filter for smoothing the signal and detected peaks based on zero-crossing. We used the count of peaks within a 10-second interval as a feature.

EEG data: For the EEG data, we extracted band power features using the biosppy package. We first applied a Butterworth high-pass filter with a cutoff frequency of 4 Hz and Butterworth low-pass filter with a cutoff frequency of 40 Hz to remove high- and low-frequency noise from the EEG signal. In addition, we used a Butterworth low-pass filter with a cutoff frequency of 40 Hz. We computed the power spectrum of the signal in each frequency band using fast Fourier Transform. We divided the signal into five frequency ranges: (1) theta (4-8 Hz), (2) alpha_low (8-10 Hz), (3) alpha_high (10-13 Hz), (4) beta (13-25 Hz), and (5) gamma (25-40 Hz), extracting band power features from each range. As the Muse S device has four electrodes, we extracted five band power features from each electrode's signal, obtaining a total of 20 features.

TEMP data: We utilized body temperature data to measure the participant's arousal state. We calculated mean, standard deviation, min, max, and slope values of the temperature data within a 10-second interval and used those as a total of 4 features.

ECG data: For the ECG data, we extracted well-known heart rate variability (HRV) features using the HeartPy package. We downloaded RR interval data from the Polar Beat application. As the validity of those values has been demonstrated [91], we utilized them directly to compute time-domain HRV features. Time-domain features include the mean of RR (Mean RR), beats per minute (BPM), the standard deviation of the RR (SDNN), the root mean square of the successive differences of RR (RMSSD), and the proportion calculated by dividing the number of interval differences of successive RR greater than 50 ms by the total number of RR (PNN50). Using a 10-second window, we obtained a total of 5 features. Additionally, we used heart rate (HR) data from the Polar Beat application. We extracted the mean and standard deviation of HR values every 10 seconds as features.

ACC data: We collected wrist acceleration data using the E4 wristband. The acceleration data was recorded on three axes (X, Y, and Z directions). We calculated mean and standard deviation values of each axis data within a 10-second interval, obtaining a total of 6 features. We calculated the magnitude of acceleration from three axis values. We used the average magnitude every 10 seconds as a feature.

4.2 Data Labeling

In this data collection study, we provided different types of stimuli with varying intensities (i.e., neutral, shouting, or swearing) to emotion workers based on the scenario protocol. As illustrated earlier, we consider two different approaches for workload labeling: i.e., (1) 'given emotional workload' conditions as labels (e.g., neutral, shouting, or swearing conditions), and (2) a worker's 'perceived emotional workload' based on self-reports as labels (e.g., low and high workload).

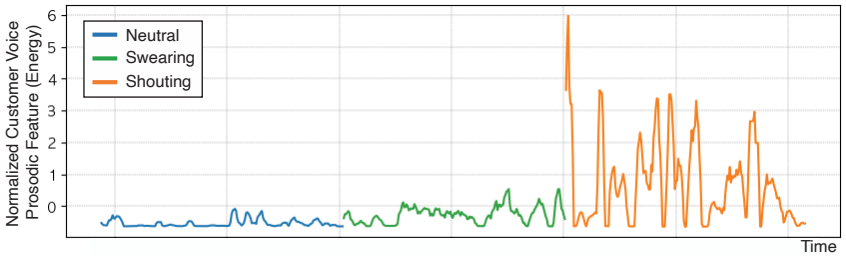

Fig. 4. An example of the normalized prosodic feature (energy) of the customer's voice. Energy, also known as volume or intensity, serves as a representation that reflects changes in the amplitude of a speech signal over time. The customer did not shout in neutral and swearing conditions. Conversely, the customer was asked to consistently shout in the shouting condition, and the customer complied well with that requirement.

4.2.1 Call-level approach for estimating given emotional workload. Based on the calls comprising the study protocol, we selected two labeling variations as follows. Firstly, we labeled the features based on the presence or absence of stimuli in the corresponding condition. Under the neutral tone condition, the customer did not raise their voices or use words that could affect the worker's emotions. In contrast, under the shouting and swearing conversation conditions, the customer raised their voices or used words that could negatively impact the emotions of the worker, thus inducing an emotional workload for the worker. Based on this, we labeled the features during neutral tone conversations as 'low-stimulus' and the features during shouting and swearing conversations as 'high-stimulus'. Secondly, we used three-class labels by considering the type of stimuli. Although both shouting and swearing conditions could induce a high emotional workload, we assume that the types of stimuli and the worker's responses differ. Therefore, we labeled the features during neutral tone conversations as 'low-stimulus', shouting condition features as 'high-stimulus (shouting)', and swearing condition features as 'high-stimulus (swearing)'.

4.2.2 Momentary approach for estimating perceived emotional workload. In the momentary approach, the workers continuously labeled their perceived emotional workload by reviewing the recorded session on a momentary basis. For example, in the shouting condition, the intensity of shouting could vary depending on the content of the conversation, and thus, individual worker may differently perceive emotional workload in different parts. To build classification models, we binarized the label (i.e., suppression) based on the self-reported data provided by the workers. The binarization threshold was set as the average of the self-reported values reported by each worker. If the value exceeded the threshold, it was labeled as 'suppressed (high-workload)'; otherwise, it was labeled as 'not suppressed (low-workload).' Furthermore, we also categorized the self-reported labels into three classes based on percentile values, representing low, medium, and high workloads. Similar to the binarization threshold, we set the thresholds at 33rd and 66th percentiles of the self-reported values. If the value was less than 33rd, it was labeled as 'low', if it fell between 33rd and 66th, it was labeled as 'medium', and if it was greater than 66th, it was labeled as 'high' workload.

4.3 Classification Pipeline

For the main analysis, we utilized a total of seven machine learning models: Decision Tree (DT) [67], Random Forest (RF) [14], AdaBoost (AB) [29], XGBoost (XGB) [15], Linear Discriminant Analysis (LDA) [7], k-Nearest Neighbor (kNN) [24], and Support Vector Machine (SVM) [48]. All data processing was performed in Python using scikit-learn. For DT, RF, and AB, the minimum number of samples required to split a node was set to 20 [95], and the decision tree was used as the base estimator for AB. For ensemble learners such as RF, AB, and XGB, the

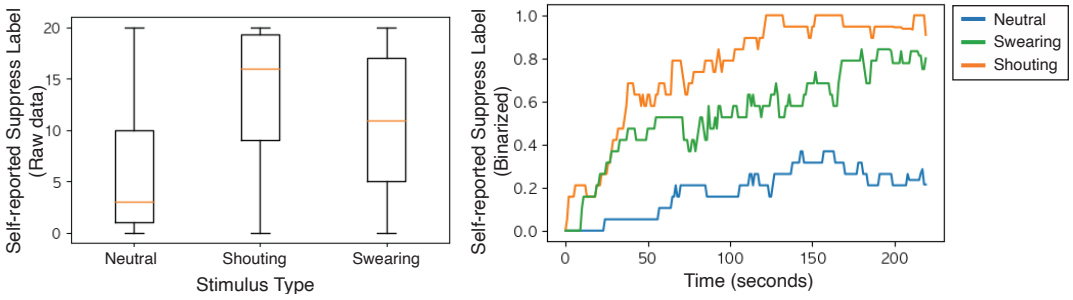

Fig. 5. Perceived emotional workload in the momentary approach. The left figure depicts a box plot of self-reported suppress raw data on a 1 to 20 scale, grouped by call conditions. The right figure illustrates the time-based trend (over 4 minutes) by averaging binarized self-report suppress data from all workers.

number of base estimators was set to 100, and XGBoost [15] was implemented using the default parameters (i.e., a learning rate of 0.1 and a max depth of 3) [21, 79] with the XGBoost package in Python.

Additionally, we employed three types of Deep Learning architectures: Fully Convolutional Network (FCN), and Multi-Layer Perceptron with LSTM (MLP-LSTM). Dzieżyc et al. [25] demonstrated that FCN performed best among the 10 DL architectures in the emotion classification task. In addition, LSTM, which is widely used in DL-based emotion recognition research [71] and suitable for learning sequential data, was included. Each signal data was used as an individual input, and details on the implementation of DL architectures were referred to existing research [25, 54].

4.3.1 Evaluation Metrics. For the binary labels in the call-level approach and momentary approach (i.e., low-stimulus: 'not suppressed' and high-stimulus: 'suppressed'), we employed accuracy, AUROC, and F1-scores. For the three-class labels (low-stimulus: 'neutral', high-stimulus: 'shouting', and 'swearing') in the call-level approach, we used accuracy and macro-averaged F1-score. Furthermore, due to potential data imbalance in the three-class labels, we reported the weighted F1-score as an evaluation metric and reported AUC-ovo (One-versus-One).

4.3.2 Validation Procedure. To validate the robustness of the models on unseen users, we used the leave-one-subject-out (LOSO), which is the method to train the model on the data of all participants except one and test the model by using excluded data and repeat it for all participants [96, 119].

5 RESULTS

The dataset from each participant consisted of three calls, with each call lasting about 4 minutes (i.e., about 12 minutes of data per participant). The average duration of a call was 4.13 minutes (SD = 0.15), and the total duration of calls from all participants was 382.05 minutes. When using 1-second windowing, the number of windows was given as 20,825. For the call-level approach with binary labeling, the number of 'low-stimulus' and 'high-stimulus' labels was 7,001 and 13,824, respectively; and with three-class labeling, the number of 'neutral', 'shouting', and 'swearing' labels was 7,001, 7,102, and 7,078, respectively. In the momentary approach with binary labeling, the number of labels for 'not suppressed (low-workload)' and 'suppressed (high-workload)' was 10,544 and 10,556, respectively; and with three classes labeling, the number of labels for 'low-workload' and 'medium-workload' and 'high-workload' was 7,723, 5,578 and 7,798, respectively.

Table 2. Evaluation of the classifiers on the binary label with two approaches: Given-Emotional-Workload (Low-stimulus vs. High-stimulus) and Perceived-Emotional-Workload (Low-workload vs. High-workload based on suppression from self-report)

Model	Given-Emotional-Workload (Low vs. High stimulus)			Perceived-Emotional-Workload (Low vs. High workload)		
	Accuracy	**F1-score**	**AUC score**	**Accuracy**	**F1-score**	**AUC score**
Decision Tree	0.78 (0.14)	0.83 (0.11)	0.77 (0.15)	0.61 (0.09)	0.60 (0.10)	0.63 (0.09)
Random Forest	0.86 (0.15)	0.90 (0.11)	0.92 (0.13)	**0.69 (0.12)**	0.68 (0.14)	0.77 (0.15)
AdaBoost	0.86 (0.17)	0.89 (0.12)	0.91 (0.15)	**0.69 (0.11)**	0.69 (0.13)	0.77 (0.15)
XGBoost	0.85 (0.17)	0.88 (0.13)	0.90 (0.15)	0.68 (0.12)	0.67 (0.14)	0.77 (0.16)
LDA	0.87 (0.13)	0.89 (0.11)	0.92 (0.12)	0.67 (0.12)	0.67 (0.15)	0.75 (0.16)
kNN	0.81 (0.06)	0.86 (0.05)	0.89 (0.06)	0.64 (0.09)	0.65 (0.10)	0.71 (0.12)
SVM	**0.87 (0.13)**	0.90 (0.10)	0.92 (0.12)	0.64 (0.11)	0.64 (0.13)	0.71 (0.15)
FCN	0.81 (0.12)	0.77 (0.14)	0.91(0.10)	0.61 (0.09)	0.58 (0.10)	0.71 (0.14)
MLP-LSTM	0.76 (0.10)	0.81 (0.12)	0.78 (0.15)	0.63 (0.14)	0.62 (0.15)	0.69 (0.16)
Majority Voting	0.66 (0.01)	0.80 (0.01)	0.50 (0.00)	0.57 (0.06)	0.43 (0.37)	0.50 (0.00)
Average	0.84 (0.14)	0.88 (0.11)	0.88 (0.12)	0.66 (0.11)	0.66 (0.13)	0.73 (0.13)

Table 3. Evaluation of the classifiers on the three-class label with two approaches: Given-Emotional-Workload (Low-stimulus vs. High-stimulus (shouting) vs. High-stimulus (swearing)) and Perceived-Emotional-Workload (Low-workload vs. Medium-workload vs. High-workload based on suppression from self-report)

Model	Given-Emotional-Workload (Low vs. High (shouting) vs. High (swearing) stimulus)			Perceived-Emotional-Workload (Low vs. Medium vs. High workload)		
	Accuracy	**F1-macro**	**AUC score**	**Accuracy**	**F1-macro**	**AUC score**
Decision Tree	0.77 (0.16)	0.76 (0.16)	0.84 (0.12)	0.42 (0.07)	0.40 (0.06)	0.57 (0.06)
Random Forest	**0.84 (0.18)**	0.84 (0.18)	0.93 (0.11)	**0.54 (0.13)**	0.47 (0.12)	0.68 (0.11)
AdaBoost	0.83 (0.19)	0.83 (0.19)	0.93 (0.11)	**0.54 (0.12)**	0.45 (0.10)	0.68 (0.11)
XGBoost	0.82 (0.22)	0.82 (0.22)	0.92 (0.12)	0.52 (0.13)	0.46 (0.12)	0.67 (0.13)
LDA	0.83 (0.19)	0.83 (0.19)	0.93 (0.11)	0.52 (0.11)	0.48 (0.10)	0.68 (0.11)
kNN	0.76 (0.07)	0.76 (0.07)	0.89 (0.05)	0.48 (0.08)	0.46 (0.08)	0.64 (0.08)
SVM	0.82 (0.18)	0.82 (0.18)	0.91 (0.12)	0.50 (0.09)	0.47 (0.07)	0.67 (0.07)
FCN	0.74 (0.12)	0.71 (0.14)	0.90 (0.07)	0.52 (0.12)	0.50 (0.12)	0.65 (0.12)
MLP-LSTM	0.70 (0.10)	0.68 (0.11)	0.87 (0.08)	0.53 (0.16)	0.55 (0.14)	0.68 (0.12)
Majority Voting	0.35 (0.01)	0.17 (0.00)	0.50 (0.00)	0.39 (0.06)	0.19 (0.02)	0.50 (0.00)
Average	0.81 (0.17)	0.81 (0.17)	0.91 (0.11)	0.50 (0.10)	0.45 (0.09)	0.66 (0.10)

5.1 RQ1: Performance of various classifiers

We performed two binary classification tasks and two three-class classification tasks based on given and perceived emotional workload different labeling approaches. Table 2 and Table 3 presents the results for each classification model. To validate the performance, we included the baseline model of majority voting, which always chooses the majority class. Each setup was run five times to report the means of the evaluation metrics.

5.1.1 Binary Labeling. For the Given-Emotional-Workload binary label, there were one call 'low-stimulus' and two calls with a 'high-stimulus', resulting in a slight imbalance in the data with durations of 4 minutes and 8 minutes, respectively. Therefore, the accuracy of the majority voting was 0.67, F1-score was 0.80, and AUC score was 0.50. The classifiers showed an average accuracy of 0.84 (SD=0.14) and an average F1-score of 0.88 (SD=0.11). The best-performing model was the SVM classifier, which had an accuracy of 0.87, approximately 0.03 higher than the majority voting, and an F1-score that was 0.10 higher. The SVM Model showed higher performance than the DL models FCN (Accuracy = 0.81) and LSTM (Accuracy = 0.76). Most of the ensemble classifiers (RF, AB, and XGB) perform better than other models. Additionally, we observed that, for the perceived emotional workload, the model's performance slightly improved when the window for physiological data was set to 10 seconds compared to setting it to 60 seconds. For the given-emotional-workload-based binary label, models using two different window sizes showed similar performance (accuracy=0.87, F1-score=0.90, AUC score=0.94 for a 60-second window, and accuracy=0.87, F1-score=0.91, AUC score=0.93 for a 10-second window). However, for the perceived-emotional-workload-based binary label, the model's performance was slightly better when extracting features from physiological data with a 10-second window (accuracy=0.64, F1-score=0.61, AUC score=0.74 for a 60-second window, and accuracy=0.67, F1-score=0.65, AUC score=0.77 for a 10-second window). Furthermore, we validated the performance even when data was not duplicated in feature extraction by setting the sliding to 0 seconds. When using the RF model as a reference, the accuracy, f1-score, and auc score were 0.87, 0.90, and 0.91, respectively, with sliding of 0 seconds, which is similar to the performance when the sliding was set to 1 second (accuracy, f1-score, and auc score were 0.86, 0.90, and 0.92, respectively). We also verified the performance even in a scenario where the data of customers (actors) included in the test set did not overlap with the data in the training set. The accuracy, F1-score, and AUC score of that model were 0.85, 0.89, and 0.90, respectively, demonstrating similar performance to the normal model.

For the Perceived-Emotional-Workload binary label, the average accuracy of majority voting was 0.57 (SD=0.06), and the f1-score was 0.43 (SD=0.37). The classifiers showed an average accuracy of 0.66 (SD=0.11) and an average F1-score of 0.66 (SD=0.13). When modeling with the same classifiers, the self-report-based labeled data showed lower performance compared to stimulus-based labeled data. The accuracy was on average 0.18 lower, and the F1-score was on average 0.22 lower. The best-performing classifiers were AB and LDA with an accuracy of 0.68.

5.1.2 Three-Class Labeling. For the Given-Emotional-Workload Three-Class label, the majority voting resulted in an average accuracy of 0.35 (SD=0.01) and F1-macro of 0.17 (SD=0.00). The best-performing Random Forest model showed an average accuracy of 0.84 (SD=0.18). Despite the increase in the number of labels, the performance was comparable to the performance of the given-emotional-workload binary model.

For the Perceived-Emotional-Workload Three-Class label, the majority voting resulted in an average accuracy of 0.39 (SD=0.06) and F1-macro of 0.19 (SD=0.02). The classifiers showed an average accuracy of 0.50 (SD=0.10) and an average F1-macro of 0.45 (SD=0.09). The best-performing model RF and AB showed an average accuracy of 0.54. Like binary labeling, perceived-emotional-workload three-class labeling showed lower performance than given-emotional-workload three-class labeling. Unlike given-emotional-workload labeling, the performance dropped compared to the perceived-emotional-workload binary model.

5.2 RQ2: Data Source Ablation Evaluation

To explore dominant features for recognizing emotional workload, we modeled various combinations of data sources and evaluated their performance. Specifically, since workers try to hide their own emotions and display the required emotions during conversations, we hypothesized that features extracted from *worker's voice data* would not help recognize emotional workload. Therefore, we compared the performance of models with and without including worker's voice features, as shown in Table 4. We used the RF classifier, which showed the best performance when aggregating a total of four (i.e., two binary and two three-class) classification tasks. In

Table 4. Evaluation of the Random Forest classifier on the binary label (High-stimulus vs. Low-stimulus) using various combinations of data sources. W.voice refers worker's voice, and C.voice refers customer's voice.

Data Combination	Random Forest (Low stimulus vs. High stimulus)		
	Accuracy	**F1-score**	**AUC score**
W.voice	0.65 (0.04)	0.78 (0.02)	0.60 (0.10)
C.voice	0.88 (0.18)	0.91 (0.12)	0.92 (0.16)
C.voice + W.voice	0.88 (0.17)	0.91 (0.12)	0.92 (0.15)
Physiological	0.75 (0.17)	0.79 (0.14)	0.76 (0.24)
Physiological + W.voice	0.73 (0.16)	0.80 (0.12)	0.77 (0.22)
C.voice + Physiological	0.86 (0.14)	0.90 (0.10)	0.93 (0.11)
C.voice + Physiological + W.voice	0.87 (0.13)	0.90 (0.10)	0.93 (0.11)
Physiological (Convenient to obtain - EDA, ACC, TEMP)	0.64 (0.15)	0.73 (0.12)	0.63 (0.23)
C.voice + W.voice + Physiological (EDA, ACC, TEMP)	0.85 (0.19)	0.89 (0.14)	0.90 (0.17)
Majority Voting	0.66 (0.01)	0.80 (0.01)	0.50 (0.00)

summary, the combinations 'customer voice' and 'customer voice+worker voice' showed the highest accuracy of 0.88 (SD=0.18) and 0.88 (SD=0.17), respectively. Conversely, when using only the worker voice, the accuracy was 0.65 (SD=0.04), and the F1-score was 0.78 (SD=0.02), showing the lowest performance comparable to the majority voting.

First, we compared the performance of models trained with classified features to examine the influence of situational cues, required emotional response, and not-required emotional response. When modeling using only the features of each classification (i.e., customer's voice features, physiological features, and worker's voice features), the accuracies were 0.88 (SD=0.18), 0.75 (SD=0.17), and 0.65 (SD=0.04), respectively. The highest performance was achieved when using only the customer's voice features, while the lowest performance was observed when using only the worker's voice features. To investigate the influence of worker's voice features on model performance, we observed the change in performance when adding 88 features from worker's voice to the following feature combinations: (1) 88 features from customer's voice, (2) 52 physiological features, and (3) 140 features from customer's voice and physiological features. Despite adding a new type of feature in all three combinations, accuracy and F1-score remained the same, indicating that worker's voice features do not contribute to improving model performance. Furthermore, we conducted modeling using only easily obtainable physiological data (i.e., EDA, ACC, TEMP) assuming in the real-world context, as well as reporting the results incorporating audio data with the obtainable physiological data. While modeling solely with easily obtainable physiological data did not yield high performance with an accuracy of 0.64 (SD=0.15), incorporating audio data in the modeling process resulted in achieving a significantly higher accuracy of 0.85 (SD=0.19).

Second, we used SHapley Additive ex-Planations (SHAP) [69] which is the framework for model agnostic interpretability to evaluate the importance of each of the 228 features in the RF Classifier. For a given feature, SHAP calculates Shapley value of each instance, which is the weighted average of marginal contributions. SHAP package visualizes Shapley value of each instance (X-axis) in a scatter plot for each feature, as in Figure 6. Feature values are color-coded, with red representing a high feature value and blue representing low feature values. For example, if the set of red dots tends to be on the right side, high values positively affect the model output, but if the set of blue dots tends to be on the right side, low values positively affect the model output.

Our result in Figure 6 shows the top 20 highly correlated features of the RF classifier, including binary call labels, using all features from 'customer's voice + physiological data + worker's voice'. The three of the top 20 features were from worker's physiological data and the others were from customer's voice. In addition to SHAP

Fig. 6. The Top 20 Features of SHapley Additive ex-Planations (SHAP): Evaluation of the importance among the total 228 features from situational cue (customer's voice), required emotion (worker's physiological data) and not-required emotion (worker's voice) in the RF classifier, including binary call labels.

value analysis, tree-based models such as RF and XGB can also utilize feature importance analysis based on mean decrease in impurity for all trees or permutation feature importance method to monitor performance changes by shuffling features. However, we note that the results of the feature importance analysis indicate how each feature is correlated to the specific classification task rather than demonstrating a causal relationship.

6 DISCUSSION

This study demonstrated that it is possible to automatically assess emotional workload in workplaces where emotional labor is required. In this section, we discuss the significance and feasibility of automatically assessing emotional workload, the strengths and weaknesses of the modeling approaches we attempted, and the ethical implications of adapting the model for capturing emotional workload in workplaces.

6.1 Theoretical Framework to Automatic Assessment of Emotional Workload

Firstly, to address the first research question, we presented a data collection and modeling method for automatically assessing emotional workload and reported the modeling results. Previous studies in the field of psychology have primarily provided theoretical frameworks for understanding the mechanisms, reasons, and consequences of emotional workload in the workplace [41, 66, 104]. Moreover, previous attempts to quantitatively measure emotional workload often relied on manual methods like self-reporting. To the best of our knowledge, however, there have been no previous attempts to automatically assess emotion regulation using data. In this study, we

collected data on factors related to emotional workload in situations where workers perform emotional labor, based on existing theoretical frameworks. One of the key distinctive features of the data collection scenarios used in our study is that we employed *interpersonal interaction* situations where we manipulated customer behavior and continuously captured a worker's workload. Through this approach, we were able to assess emotional workload when various levels and types of situational cues were given to workers. By leveraging the collected multimodal dataset and employing various machine learning algorithms, we proposed models for recognizing emotional workload, demonstrating that the emotional workload can be assessed automatically. Furthermore, our approach allows for real-time assessment of a worker's emotional workload, surpassing the limitations of post-event (retrospective) measures like self-report. From a retrospective annotation perspective, it also has the advantage of reducing the worker's burden (time and effort), as they do not need to additionally report their state.

As seen in Figure 1, emotion regulation has short-term effects of increasing worker's stress and reducing self-control capacity. If emotion regulation persists over the long term, the worker may find themselves in states of ego depletion and emotion exhaustion. From the perspective of a worker's well-being, this study enables real-time monitoring of a worker's emotional workload so that workers can self-track how much of their capacity they are using for emotion regulation. Furthermore, it can prevent workers from being exposed to high emotional workloads for extended periods, thus mitigating the risk of emotion exhaustion or ego depletion (e.g., nudging users to have short breaks after experiencing high emotional workloads). While there may be some correlation between emotion regulation and stress/emotion, the key feature of this study is not to recognize the worker's state after mental health issues like stress or emotion exhaustion have already occurred but to monitor how much the worker's capacity is decreasing preventively. Additionally, this research has applications in terms of worker productivity, job scheduling, performance evaluation, and customer interaction strategies. However, it requires careful ethical considerations, as discussed later. Many studies that measure cognitive load aim to consider an individual's information processing capacity for job planning or optimizing interfaces. Similarly, allowing a worker's mental state to deteriorate irreversibly can negatively impact both their well-being and performance. Therefore, it is necessary to take into account the current emotional workload of workers for efficient and safe job scheduling. Additionally, environmental support to maintain emotional workload at reasonable levels is also required.

We can also consider this issue in relation to performance evaluation. Traditionally, worker performance evaluations have relied solely on customer ratings after the service has been provided. This can be a subjective assessment from the customer's perspective. In this context, the measured workload can serve as a source that captures the context of interactions between customers and workers during the service. Accordingly, performance evaluations can be carried out more objectively than with the traditional system.

6.2 Emotional Workload Modeling with a Multimodal Dataset

6.2.1 Key Factors for Recognizing Emotional Workload. We explored the importance of different data sources (e.g., situational cues - customer's voice, required emotional response - worker's voice, not-required emotional response - worker's physiological response) in recognizing emotional workload. To achieve this, we compared the results of models using various data combinations based on the RF classifier. Interestingly, the accuracy of the model using only customer's voice data was 0.88 (SD=0.18), and the model using only worker's physiological response data had an accuracy of 0.75 (SD=0.17), both outperforming the model using only worker's voice data. Additionally, the inclusion of worker's voice features did not improve the performance compared to models using customer's voice, physiological response, or customer's voice, physiological response data alone. We explain that the reason for the lowest performance of worker's voice data is that the worker's voice represents the result of acting in accordance with display rules. In call centers, workers receive evaluations of call quality from both customers and company managers, and they undergo training to achieve high-quality call interactions, including

maintaining a specific tone of voice [58, 115]. Therefore, using a worker's voice alone makes it difficult to estimate the emotional workload that workers experience. This result contrasts with previous research findings that detect stress and emotions using human voice [59, 68].

Furthermore, according to the existing theoretical framework for emotion regulation, workers are known to be significantly influenced by external stimuli [37]. For example, in situations where display rules are imposed and high-intensity situational cues are present, workers are required to perform higher emotional regulation. Similarly, our study's results also demonstrated that the model using only the worker's physiological response data (accuracy=0.75, SD=0.17) was outperformed by the model that incorporated a customer's voice data (accuracy=0.86, SD=0.14). Through this research, we confirmed that it is essential to utilize situational cues and not-required emotional responses for recognizing emotional workload, while required emotional responses did not provide significant assistance.

6.2.2 Differences in Labeling Approaches: Given- vs. Perceived-Emotional-Workload. In this study, we distinguished emotional workload into two categories: objectively *given emotional workload* and subjectively *perceived emotional workload* by the worker. Accordingly, we employed two different labeling methods for data. As a result, we confirmed that when labeling was based on the given emotional workload, the model's accuracy and F1-score were higher than when labeling was based on perceived emotional workload. This outcome aligns with similar findings in previous research, where using self-report data to detect stress or emotions resulted in lower performance [80, 107]. We provide several possible explanations for these results.

We analyzed the correspondence between stimulus-based labeling data and self-reported labeling data to understand the disparities in performance between the given emotional workload model and the perceived emotional workload model. First, data exhibiting identical trends in both labels (i.e., data labeled as either 0 or 1 in both label sets) accounted for 73.62% of the total dataset. Conversely, data labeled differently comprised 26.38% of the dataset, with 5.94% recorded as having a stimulus-based label of 0 but a self-report-based label of 1, and 20.44% recorded as having a stimulus-based label of 1 but a self-report-based label of 0. In this way, a significant portion of label disparities was attributed to workers' reporting that they were not regulating their emotions even in situations involving stimuli such as shouting or criticism. However, modeling results utilizing data collected from only workers (workers' physiological and voice data) without customers' data revealed that the accuracy of the RF model utilizing self-reported labels was 0.58 (SD=0.11), lower than when utilizing stimulus-based labels (0.70 (SD=0.12)). Although workers perceived and reported they did not make efforts to regulate their emotions in response to stimuli, the results may indicate their physiological responses were more aligned with the given-emotional workload. We provide several possible explanations for these results.

The first explanation is that there can be individual differences in how humans report their emotional status depending on their emotional granularity and emotional complexity. Emotional granularity refers to an individual's unique ability to discern and express similar emotions in a detailed and nuanced manner [98]. Consequently, there are variations among individuals in the breadth and distinctiveness of their emotional experiences, a concept often referred to as emotional complexity [57]. Emotional complexity has been known to be associated with private self-consciousness, openness to experience, empathic tendencies, cognitive complexity, the ability to distinguish among named emotions, the range of emotions experienced daily, and interpersonal adaptability. This means that, even when presented with situational cues of the same intensity level, individuals may perceive emotions differently based on their personal emotional predispositions and experiences. Furthermore, the ability to differentiate one's own emotions can influence the self-reported results of one's emotional state. For example, when dealing with a customer who is yelling, depending on the worker's unique emotional range, one person may experience very negative emotions while another may experience relatively less negative emotions as in Figure 7. During the process of reporting their perceived emotions, some individuals may finely differentiate their subtle emotional nuances and report them on a detailed scale, while others may express their emotions in a

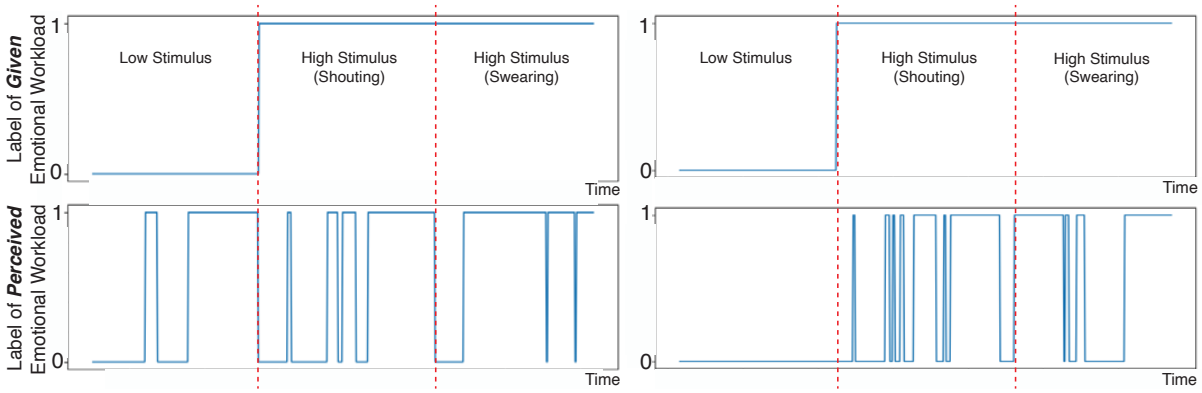

Fig. 7. Individual differences in self-reported suppress. The two graphs on the left are the given and perceived emotional workload for one subject, and the two graphs on the right are the given and perceived emotional workload for another subject. Even when presented with situational cues of the same intensity level, individuals may perceive emotional suppression differently based on emotional predispositions and prior experiences. In this experiment, even in low-stimulus conditions, some workers responded that they experienced suppression, while others reported that they did not.

binary fashion. These individual differences can be factors that potentially impair the performance of generalized models.

Another explanation suggests that there may be differences between the rate of change in the situational cue, which is the customer's voice variation, and the rate of change in physiological responses. As seen in Table 4, even in a model that uses only customer voice features, the accuracy was around 0.88 (SD=0.18), indicating that workers were sensitive to the influence of situational cues on their emotion regulation. Indeed, according to the result of observing the self-reporting process of the workers, it was evident that the workers' level of suppression in their responses reflected the influence of the customer's voice in real-time. Unlike situations that typically elicit a sustained level of stress, such as immersing one's hand in cold water or performing mathematical calculations, our data collection study involved situational cues where the stimulus magnitude changed in real-time during the conversation. For example, in the shouting customer condition, the customer did not continuously shout from the beginning to the end of the call but rather interjected frequently with loud outbursts during the conversation. In such cases, the extent of suppression reported by workers could also have changed in real-time in response to the situational cue. However, typically, physiological responses such as HR or EDA take time to return to their original state after a stress-induced change, and this recovery time can vary among individuals [93]. Therefore, even if a worker reports less suppression in response to a reduced situational cue, their physiological response may still indicate that they are still highly engaged in emotion regulation.

Nevertheless, physiological responses can still serve as a crucial feature for workload measurement on their own. As physiological responses such as HR, EDA, and EEG are objective numerical metrics that can identify a person's physical condition, it has been widely used as features not only in the field of affective computing for detecting stress and emotions but also in the measurement of cognitive workload [113]. When modeling with workers' physiological data alone, higher performance was observed when using stimulus-based labels compared to self-report labels. It may indicate that, in situations where psychological constraints are imposed through display rules, workers find it difficult to perceive the emotional and psychological changes they are experiencing. Exposed to continuous negative stimuli without accurately recognizing their own state, workers may become emotionally exhausted, leading to more serious consequences such as ego-depletion or burnout. Therefore, designing intervention systems such as temporarily excluding workers from tasks or recommending

breaks when third-party observation indicates sustained regulation beyond a certain level could be beneficial. Finding appropriate timing for emotional workers to recover can aid in their recuperation.

Moreover, our research demonstrated the need for improved models assessing emotion regulation from a first-person perspective, such as the perceived-emotional-workload model. For a more accurate emotional workload model using self-report-based labels, the following additional studies are needed. Firstly, from the perspective of emotional granularity, workers' characteristics or abilities in reporting their emotions should be included in the modeling process [8, 50]. Individuals may react differently to external stimuli based on factors such as career, experience, and personality. This research did not account for such individual traits, and the absence of personalized characteristics might have significantly impacted the perceived-emotional-workload model's low performance with the workers from various distributions, although it enhances the model's generalizability. Second, future works have to consider the subject's physical characteristics such as the speed at which the body responds to various types of stimuli. For instance, the body's response time may vary depending on the type of stimulus (visual, auditory), and individual differences may also exist. Integrating these factors could potentially lead to the development of more effective personalized models.

6.3 Ethical Implications

Monitoring workers' emotional workloads in real-time can be highly beneficial for maintaining their well-being because workers can better self-track their workloads. However, when data collection is performed in a real-world context, it can also be a double-edged sword, raising several potential ethical issues such as violating workers' autonomy through workplace surveillance [118], data misuse/abuse [76], and infringement on customer privacy.

Firstly, even if emotion AI can be utilized to improve work productivity and well-being with consent, the act of being monitored may be perceived as surveillance by workers. Assessing emotional workload per se can become a source of additional negative emotions or stress for workers, although call quality monitoring is a common industry practice [51]. Furthermore, the results evaluating emotional workload might enable objective performance evaluations so it can be perceived as another burden to workers. From the worker's perspective, it can be regarded as the emergence of another 'computerized' metric for workers' performance evaluation, commonly referred to as Digital Taylorism [3, 53, 106].

Also, during the data collection phase for modeling emotion AI, the privacy of both workers and customers can be compromised. Our dataset includes audio data containing conversations between customers and workers, which may contain sensitive information about the individuals involved. Therefore, collecting audio data in situations where actual important personal information such as finance or insurance industry contexts poses significant risks in terms of privacy, unlike situations based on lab-created scenarios. Additionally, The emotion AI may threaten worker privacy as well. workers consider their emotions to be private information in the workplace, and the utilization of AI for emotion recognition may be perceived as an infringement on their privacy [92]. While emotional labor work mandates 'display rules' (what is being observed), workers also deserve the right to keep their emotions private. Various types of input data are utilized for emotion AI techniques, including biological sensor data, facial micro-expressions, physiological and speech signals, and text semantics. Among them, if data (e.g., bvp, eeg) unintentionally allows inferring states (e.g., Health status or cognitive-behavioral abilities) other than worker emotions, the scope of invading workers' privacy broadens, increasing the risk. Thus, while diverse data collection enables high model performance, it also poses significant risks to the privacy of both workers and customers if misused or exploited.

Nevertheless, we emphasize the necessity of assessing emotional workload in real workplaces to prevent workers from experiencing irreversible mental or psychological states for personal and organizational purposes. For that, emotion recognition technology must be accompanied by various institutional systems. Firstly, workers should know what information is being collected and understand accurately the trade-off relationship between

workplace surveillance and psychological well-being [92]. Data collectors and workers should transparently share all details related to data collection, including data collection items, retention periods, and usage methods. Secondly, laws and regulations need to be established at the national level to protect workers when using emotional workload assessment in actual workplaces. For example, when managers intend to use workers' emotional workload data, laws should be enacted to restrict access to low-level raw data that enables additional analysis or interpretation of a worker's state. Restricting the usage scope can help workers feel they are not closely monitored by the company or managers. Thirdly, encrypting data can help with data privacy. With the proliferation of monitoring using electronic devices like smart home systems, user data privacy concerns are rising. While mass data collection can bring social benefits, privacy concerns often arise, and using encryption technologies like homomorphic encryption can enable freer sharing of data with external parties. Additionally, there are things to consider when using the system in this study directly in the real world. The sensors we used for data collection, such as the muse (headband), E4 band (wristband), and polar h10 (chest band), all require body contact for accurate data measurement. Therefore, workers need to be adequately informed beforehand about discomfort due to device wearing and the potential limitation of movement over prolonged data collection periods.

7 LIMITATIONS

The method proposed in our research for assessing emotional workload using multimodal sensing and machine learning contributes to the well-being and productivity of workers. However, there are several limitations to this approach.

First of all, our data collection protocol has some limitations because it was not conducted in a real-world environment [87]. Typically, workers do not frequently engage in consecutive calls with the same customer. However, in this experiment, the scenario involved three consecutive calls with the same customer. Additionally, even though we made efforts to replicate a work environment as closely as possible through scenario specification, the absence of systems allowing access to customer information might have made the experimental environment feel unfamiliar to the workers. Nevertheless, in the post-survey, workers responded they did not feel uncomfortable with consecutive calls with the same person, and they expressed a high level of immersion due to the similarity of the customer's dialogue and tone to real interactions.

Secondly, there are limitations to applying our research methodology to the real world. While our model showed good performance based on data collected from audio data and easily wearable sensors, the utilization of audio data may be challenging if the data contains overly private information. Moreover, requiring intrusive devices for higher performance may impose burdens on workers. Additionally, limitations may arise when applying the completed model to the real world. To achieve high performance when encountering completely new users, a certain amount of data collection is necessary for normalizing individual data. Before a sufficient amount of data has been collected, the model may classify without adequately considering individual differences, potentially resulting in somewhat lower accuracy. Nevertheless, as more data accumulates over time, accuracy is expected to approach our benchmarking results.

Lastly, there are limitations related to the modeling process. Firstly, our model did not consider linguistic-related stimulus features that may occur during calls between customers and workers. For instance, a customer might not raise their voice but may use aggressive words under the swearing condition. In such cases, contrary to our intention, the model may perceive the situation as a low-stimulus condition in the binary labeling session. In the case of three-class classification, a situation with low stimuli but high perceived workload can be classified as a swearing condition, as intended. For these reasons, the accuracy of binary and three-class classification showed similar performance in our study, unlike the results of previous studies that generally show higher performance in binary classification than in three-class classification. Furthermore, this study did not account for individual

differences in the ability or characteristics of workers to report emotions. Given these limitations, we anticipate future research to address features related to linguistic stimuli and individual differences in expressing emotions based on display rules.

8 CONCLUSION

In this study, we proposed machine learning models capable of automatically assessing workers' emotional workload in the context of emotional labor by utilizing multimodal sensor data. We designed the data collection study based on a previously established theoretical framework in the field of psychology and collected the multimodal dataset using wearable sensors. By providing the results of benchmarking, we verified that machine learning models can assess the emotional workload under the contact center scenario where display rules are provided. Additionally, we explored the dominant features of emotional workload assessment. As a result, the SVM classifier-based model exhibited the highest accuracy, reaching 0.88 not only in binary classification but also in three-class classification. This research provided a definition of emotional workload and demonstrated that emotional workload can be assessed with high accuracy through multimodal sensing.

ACKNOWLEDGMENTS

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). Also, this research was supported by the Chung-Ang University Research Grants in 2023.

REFERENCES

[1] Shazia Afzal and Peter Robinson. 2009. Natural affect data—Collection & annotation in a learning context. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 1–7.

[2] Mehmet Berkehan Akçay and Kaya Oğuz. 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication* 116 (2020), 56–76.

[3] Moritz Altenried. 2020. The Platform as Factory: Crowdwork and the Hidden Labour Behind Artificial Intelligence. *Capital & Class* 44, 2 (2020), 145–158.

[4] David M Amodio, Leah R Zinner, and Eddie Harmon-Jones. 2007. Social psychological methods of emotion elicitation. *Handbook of emotion elicitation and assessment* 91 (2007), 91–105.

[5] M Dilli Babu, DV JeevithaShree, Gowdham Prabhakar, Kamal Preet Singh Saluja, Abhay Pashilkar, and Pradipta Biswas. 2019. Estimating pilots' cognitive load from ocular parameters through simulation and in-flight studies. *Journal of Eye Movement Research* 12, 3 (2019).

[6] Jo-Anne Bachorowski and Ellen B Braaten. 1994. Emotional intensity: Measurement and theoretical implications. *Personality and individual differences* 17, 2 (1994), 191–199.

[7] Suresh Balakrishnama and Aravind Ganapathiraju. 1998. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing* 18, 1998 (1998), 1–8.

[8] Lisa Feldman Barrett and Ajay B Satpute. 2019. Historical pitfalls and new directions in the neuroscience of emotion. *Neuroscience letters* 693 (2019), 9–18.

[9] Roy F Baumeister, Ellen Bratslavsky, Mark Muraven, and Dianne M Tice. 1998. Ego depletion: Is the active self a limited resource? *Journal of personality and social psychology* 74, 5 (1998), 1252.

[10] Roy F Baumeister, Kathleen D Vohs, and Dianne M Tice. 2007. The strength model of self-control. *Current directions in psychological science* 16, 6 (2007), 351–355.

[11] Daniel J Beal and John P Trougakos. 2013. Episodic intrapersonal emotion regulation: Or, dealing with life as it happens. In *Emotional labor in the 21st century*. Routledge, 51–76.

[12] 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.

[13] Patricia J Bota, Chen Wang, Ana LN Fred, and Hugo Plácido Da Silva. 2019. A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. *IEEE Access* 7 (2019), 140990–141020.

[14] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[15] 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.

[16] 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).

[17] Youngjun Cho. 2021. Rethinking eye-blink: Assessing task difficulty through physiological representation of spontaneous blinking. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–12.

[18] George E Cooper and Robert P Harper. 1969. *The use of pilot rating in the evaluation of aircraft handling qualities*. National Aeronautics and Space Administration.

[19] Stéphane Coté. 2005. A social interaction model of the effects of emotion regulation on work strain. *Academy of management review* 30, 3 (2005), 509–530.

[20] Richard J Davidson. 1998. Affective style and affective disorders: Perspectives from affective neuroscience. *Cognition & emotion* 12, 3 (1998), 307–330.

[21] 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.

[22] James M Diefendorff and Robin H Gosserand. 2003. Understanding the emotional labor process: A control theory perspective. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 24, 8 (2003), 945–959.

[23] James M Diefendorff and Erin M Richard. 2008. Not all emotional display rules are created equal: Distinguishing between prescriptive and contextual display rules. *Research companion to emotion in organizations* (2008), 316–334.

[24] Richard O Duda, Peter E Hart, et al. 1973. *Pattern classification and scene analysis*. Vol. 3. Wiley New York.

[25] Maciej Dzieżyc, Martin Gjoreski, Przemysław Kazienko, Stanisław Saganowski, and Matjaž Gams. 2020. Can we ditch feature engineering? end-to-end deep learning for affect recognition from physiological sensor data. *Sensors* 20, 22 (2020), 6535.

[26] Sidney D'Mello and Art Graesser. 2012. Dynamics of affective states during complex learning. *Learning and Instruction* 22, 2 (2012), 145–157.

[27] Sidney K D'Mello. 2015. On the influence of an iterative affect annotation approach on inter-observer and self-observer reliability. *IEEE Transactions on Affective Computing* 7, 2 (2015), 136–149.

[28] F Thomas Eggemeier. 1988. Properties of workload assessment techniques. In *Advances in psychology*. Vol. 52. Elsevier, 41–62.

[29] Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 55, 1 (1997), 119–139.

[30] Lex Fridman, Bryan Reimer, Bruce Mehler, and William T Freeman. 2018. Cognitive load estimation in the wild. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–9.

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

[32] Junling Gao, Hang Kin Leung, Jicong Fan, Bonnie Wai Yan Wu, and Hin Hung Sik. 2022. The neurophysiology of the intervention strategies of Awareness Training Program on emotion regulation. *Frontiers in Psychology* 13 (2022), 891656.

[33] Nadia Garnefski, Vivian Kraaij, and Philip Spinhoven. 2001. Negative life events, cognitive emotion regulation and emotional problems. *Personality and Individual differences* 30, 8 (2001), 1311–1327.

[34] Catherine R Glenn, Terry D Blumenthal, E David Klonsky, and Greg Hajcak. 2011. Emotional reactivity in nonsuicidal self-injury: Divergence between self-report and startle measures. *International Journal of Psychophysiology* 80, 2 (2011), 166–170.

[35] Lori Sideman Goldberg and Alicia A Grandey. 2007. Display rules versus display autonomy: emotion regulation, emotional exhaustion, and task performance in a call center simulation. *Journal of occupational health psychology* 12, 3 (2007), 301.

[36] Daniel Gopher and Rolf Braune. 1984. On the psychophysics of workload: Why bother with subjective measures? *Human factors* 26, 5 (1984), 519–532.

[37] 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.

[38] Alicia A Grandey, David N Dickter, and Hock-Peng Sin. 2004. The customer is not always right: Customer aggression and emotion regulation of service employees. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior* 25, 3 (2004), 397–418.

[39] Kim L Gratz and Lizabeth Roemer. 2004. Multidimensional assessment of emotion regulation and dysregulation: Development, factor structure, and initial validation of the difficulties in emotion regulation scale. *Journal of psychopathology and behavioral assessment* 26 (2004), 41–54.

[40] Alberto Greco, Gaetano Valenza, Antonio Lanata, Enzo Pasquale Scilingo, and Luca Citi. 2015. cvxEDA: A convex optimization approach to electrodermal activity processing. *IEEE Transactions on Biomedical Engineering* 63, 4 (2015), 797–804.

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

[42] James J Gross. 2015. Emotion regulation: Current status and future prospects. *Psychological inquiry* 26, 1 (2015), 1–26.
[43] James J Gross and Oliver P John. 2003. Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *Journal of personality and social psychology* 85, 2 (2003), 348.
[44] Andreas Haag, Silke Goronzy, Peter Schaich, and Jason Williams. 2004. Emotion recognition using bio-sensors: First steps towards an automatic system. In *Tutorial and research workshop on affective dialogue systems*. Springer, 36–48.
[45] Eija Haapalainen, SeungJun Kim, Jodi F Forlizzi, and Anind K Dey. 2010. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. 301–310.
[46] Lloyd C Harris. 2013. Service employees and customer phone rage: An empirical analysis. *European Journal of Marketing* 47, 3/4 (2013), 463–484.
[47] Sandra G Hart. 1986. NASA task load index (TLX). (1986).
[48] 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.
[49] 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.
[50] Katie Hoemann, Zulqarnain Khan, Mallory J Feldman, Catie Nielson, Madeleine Devlin, Jennifer Dy, Lisa Feldman Barrett, Jolie B Wormwood, and Karen S Quigley. 2020. Context-aware experience sampling reveals the scale of variation in affective experience. *Scientific reports* 10, 1 (2020), 12459.
[51] David Holman, Claire Chissick, and Peter Totterdell. 2002. The effects of performance monitoring on emotional labor and well-being in call centers. *Motivation and Emotion* 26 (2002), 57–81.
[52] 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*. 493–504.
[53] John Howard. 2022. Algorithms and the future of work. *American Journal of Industrial Medicine* 65, 12 (2022), 943–952.
[54] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. 2019. Deep learning for time series classification: a review. *Data mining and knowledge discovery* 33, 4 (2019), 917–963.
[55] Kyung-Sook Jeong, Su-Jeong Choi, Myeong-Ok Park, and Yan Li. 2015. 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 Journal of Business Administration* 28, 2 (2015).
[56] Rachel E Jones, Ellen W Leen-Feldner, Bunmi O Olatunji, Laura E Reardon, and Erin Hawks. 2009. Psychometric properties of the Affect Intensity and Reactivity Measure adapted for Youth (AIR–Y). *Psychological Assessment* 21, 2 (2009), 162.
[57] Sun-Mee Kang and Phillip R Shaver. 2004. Individual differences in emotional complexity: Their psychological implications. *Journal of personality* 72, 4 (2004), 687–726.
[58] Betül Karakus and Galip Aydin. 2016. Call center performance evaluation using big data analytics. In *2016 International Symposium on Networks, Computers and Communications (ISNCC)*. IEEE, 1–6.
[59] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. 2019. Speech emotion recognition using deep learning techniques: A review. *IEEE Access* 7 (2019), 117327–117345.
[60] M Asif Khawaja, Fang Chen, and Nadine Marcus. 2014. Measuring cognitive load using linguistic features: implications for usability evaluation and adaptive interaction design. *International Journal of Human-Computer Interaction* 30, 5 (2014), 343–368.
[61] Anusha Koduru, Hima Bindu Valiveti, and Anil Kumar Budati. 2020. Feature extraction algorithms to improve the speech emotion recognition rate. *International Journal of Speech Technology* 23, 1 (2020), 45–55.
[62] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2011), 18–31.
[63] Arthur F Kramer, Erik J Sirevaag, and Rolf Braune. 1987. A psychophysiological assessment of operator workload during simulated flight missions. *Human factors* 29, 2 (1987), 145–160.
[64] Sylvia D Kreibig. 2010. Autonomic nervous system activity in emotion: A review. *Biological psychology* 84, 3 (2010), 394–421.
[65] Eva G Krumhuber, Arvid Kappas, and Antony SR Manstead. 2013. Effects of dynamic aspects of facial expressions: A review. *Emotion Review* 5, 1 (2013), 41–46.
[66] Sandra A Lawrence, Ashlea C Troth, Peter J Jordan, and Amy L Collins. 2011. A review of emotion regulation and development of a framework for emotion regulation in the workplace. *The role of individual differences in occupational stress and well being* 9 (2011), 197–263.
[67] Wei-Yin Loh. 2011. Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery* 1, 1 (2011), 14–23.

[68] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T Chittaranjan, Andrew T Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. 2012. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. 351–360.

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

[70] Robert J Lysaght, Susan G Hill, AO Dick, Brian D Plamondon, Paul M Linton, Walter W Wierwille, Allen L Zaklad, AC Bittner, and Robert J Wherry. 1989. *Operator workload: Comprehensive review and evaluation of operator workload methodologies*. NTIS (Reprod.).

[71] M Maithri, U Raghavendra, Anjan Gudigar, Jyothi Samanth, Prabal Datta Barua, Murugappan Murugappan, Yashas Chakole, and U Rajendra Acharya. 2022. Automated emotion recognition: Current trends and future perspectives. *Computer methods and programs in biomedicine* 215 (2022), 106646.

[72] Sven L Mattys and Lukas Wiget. 2011. Effects of cognitive load on speech recognition. *Journal of memory and Language* 65, 2 (2011), 145–160.

[73] Iris B Mauss, Robert W Levenson, Loren McCarter, Frank H Wilhelm, and James J Gross. 2005. The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion* 5, 2 (2005), 175.

[74] Iris B Mauss, Amanda J Shallcross, Allison S Troy, Oliver P John, Emilio Ferrer, Frank H Wilhelm, and James J Gross. 2011. Don't hide your happiness! Positive emotion dissociation, social connectedness, and psychological functioning. *Journal of personality and social psychology* 100, 4 (2011), 738.

[75] Janet R McColl-Kennedy, Paul G Patterson, Amy K Smith, and Michael K Brady. 2009. Customer rage episodes: emotions, expressions and behaviors. *Journal of Retailing* 85, 2 (2009), 222–237.

[76] Andrew McStay. 2020. Emotional AI, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy. *Big Data & Society* 7, 1 (2020), 2053951720904386.

[77] Pavle Mijović, Miloš Milovanović, Ivan Gligorijević, Vanja Ković, Ivana Živanović-Mačužić, and Bogdan Mijović. [n. d.]. Investigating brain dynamics in industrial environment–integrating mobile EEG and kinect for cognitive state detection of a worker. In *Augmented Cognition. Neurocognition and Machine Learning: 11th International Conference, AC 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings, Part I 11*.

[78] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. 2018. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing* 12, 2 (2018), 479–493.

[79] Shayan Mirjafari, Kizito Masaba, Ted Grover, Weichen Wang, Pino Audia, Andrew T Campbell, Nitesh V Chawla, Vedant Das Swain, Munmun De Choudhury, Anind K Dey, et al. 2019. Differentiating higher and lower job performers in the workplace using mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–24.

[80] 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.

[81] Varun Mishra, Sougata Sen, Grace Chen, Tian Hao, Jeffrey Rogers, Ching-Hua Chen, and David Kotz. 2020. Evaluating the reproducibility of physiological stress detection models. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 4, 4 (2020), 1–29.

[82] Sebastian C Müller and Thomas Fritz. 2015. Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, Vol. 1. IEEE, 688–699.

[83] Delphine Nelis, Jordi Quoidbach, Michel Hansenne, and Moïra Mikolajczak. 2011. Measuring individual differences in emotion regulation: The emotion regulation profile-revised (ERP-R). *Psychologica belgica* 51, 1 (2011).

[84] Fred GWC Paas, Jeroen JG Van Merriënboer, and Jos J Adam. 1994. Measurement of cognitive load in instructional research. *Perceptual and motor skills* 79, 1 (1994), 419–430.

[85] Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. 2020. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data* 7, 1 (2020), 293.

[86] KFV Phillips and MJ Power. 2007. A new self-report measure of emotion regulation in adolescents: The Regulation of Emotions Questionnaire. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice* 14, 2 (2007), 145–156.

[87] Rosalind W Picard. 2016. Automating the recognition of stress and emotion: From lab to real-world impact. *IEEE MultiMedia* 23, 3 (2016), 3–7.

[88] Jordi Quoidbach, Deplphine Nelis, Moïra Mikolajczak, and Michel Hansenne. 2007. Development and validation of a typical performance Emotional Regulation Profile (ERP-Q). In *Annual meeting of the Belgian Association for Psychological Science*.

[89] Anat Rafaeli and Robert I Sutton. 1987. Expression of emotion as part of the work role. *Academy of management review* 12, 1 (1987), 23–37.

[90] Gary B Reid and Thomas E Nygren. 1988. The subjective workload assessment technique: A scaling procedure for measuring mental workload. In *Advances in psychology*. Vol. 52. Elsevier, 185–218.

[91] Polar Research and Technology. 2019. Polar H10 Heart Rate Sensor System.

[92] 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*. 1–20.

[93] Derek Roger and John Jamieson. 1988. Individual differences in delayed heart-rate recovery following stress: The role of extraversion, neuroticism and emotional control. *Personality and Individual differences* 9, 4 (1988), 721–726.

[94] Florian Schaule, Jan Ole Johanssen, Bernd Bruegge, and Vivian Loftness. 2018. 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 (2018), 1–20.

[95] 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*. 400–408.

[96] Philip Schmidt, Attila Reiss, Robert Dürichen, and Kristof Van Laerhoven. 2019. Wearable-based affect recognition—A review. *Sensors* 19, 19 (2019), 4079.

[97] Björn Schuller and Anton Batliner. 2013. *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons.

[98] Katharine E Smidt and Michael K Suvak. 2015. A brief, but nuanced, review of emotional granularity and emotion differentiation research. *Current Opinion in Psychology* 3 (2015), 48–51.

[99] Wally Smith, Greg Wadley, Sarah Webber, Benjamin Tag, Vassilis Kostakos, Peter Koval, and James J Gross. 2022. Digital emotion regulation in everyday life. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

[100] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2011. A multimodal database for affect recognition and implicit tagging. *IEEE transactions on affective computing* 3, 1 (2011), 42–55.

[101] Jiraporn Surachartkumtonkun, Janet R McColl-Kennedy, and Paul G Patterson. 2015. Unpacking customer rage elicitation: A dynamic model. *Journal of Service Research* 18, 2 (2015), 177–192.

[102] John Sweller. 2011. Cognitive load theory. In *Psychology of learning and motivation*. Vol. 55. Elsevier, 37–76.

[103] David L Tobin, Kenneth A Holroyd, Russ V Reynolds, and Joan K Wigal. 1989. The hierarchical factor structure of the Coping Strategies Inventory. *Cognitive therapy and research* 13 (1989), 343–361.

[104] Ashlea C Troth, Sandra A Lawrence, Peter J Jordan, and Neal M Ashkanasy. 2018. Interpersonal emotion regulation in the workplace: A conceptual and operational review and future research agenda. *International Journal of Management Reviews* 20, 2 (2018), 523–543.

[105] Terumi Umematsu, Akane Sano, Sara Taylor, Masanori Tsujikawa, and Rosalind W Picard. 2020. 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)*. IEEE, 5953–5957.

[106] J Vázquez and M García. 2011. From Taylorism to neo Taylorism: a 100 year journey in human resource management. *Int Rev Public Nonprofit Mark Madrid* 8, 2 (2011), 111–130.

[107] Gideon Vos, Kelly Trinh, Zoltan Sarnyai, and Mostafa Rahimi Azghadi. 2023. Generalizable machine learning for stress monitoring from wearable devices: a systematic literature review. *International Journal of Medical Informatics* (2023), 105026.

[108] Greg Wadley, Wally Smith, Peter Koval, and James J Gross. 2020. Digital emotion regulation. *Current Directions in Psychological Science* 29, 4 (2020), 412–418.

[109] Johannes Wagner, Jonghwa Kim, and Elisabeth André. 2005. From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In *2005 IEEE international conference on multimedia and expo*. IEEE, 940–943.

[110] Thomas L Webb, Eleanor Miles, and Paschal Sheeran. 2012. Dealing with feeling: a meta-analysis of the effectiveness of strategies derived from the process model of emotion regulation. *Psychological bulletin* 138, 4 (2012), 775.

[111] Christopher D Wickens. 2020. Processing resources and attention. In *Multiple task performance*. CRC Press, 3–34.

[112] Christopher D Wickens, William S Helton, Justin G Hollands, and Simon Banbury. 2021. *Engineering psychology and human performance*. Routledge.

[113] Glenn F Wilson. 2002. An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *The International Journal of Aviation Psychology* 12, 1 (2002), 3–18.

[114] Bo Yin, Fang Chen, Natalie Ruiz, and Eliathamby Ambikairajah. 2008. Speech-based cognitive load monitoring system. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2041–2044.

[115] 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.

[116] Rachel L Zelkowitz and David A Cole. 2016. Measures of emotion reactivity and emotion regulation: Convergent and discriminant validity. *Personality and Individual Differences* 102 (2016), 123–132.

[117] Jia-Lin Zhao, Xu-Hong Li, and John Shields. 2019. Managing job burnout: The effects of emotion-regulation ability, emotional labor, and positive and negative affect at work. *International Journal of Stress Management* 26, 3 (2019), 315.

[118] Kathryn Zickuhr. 2021. Workplace surveillance is becoming the new normal for US workers. *Washington Center for Equitable Growth. https://equitablegrowth. org/research-paper/workplace-surveillance-is-becomingthe-new-normal-for-us-workers/. Institute for Research on Labor and Employment University of California, Berkeley* 2521 (2021), 94720–5555.

[119] 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*. 1–14.