EmpathicSchool: A multimodal dataset for real-time facial expressions and physiological data analysis under different stress conditions

Majid Hosseini¹, Fahad Sohrab², Raju Gottumukkala¹,⁴, Ravi Teja Bhupatiraju¹, Satya Katragadda¹, Jenni Raitoharju³, Alexandros Iosifidis⁴, and Moncef Gabbouj²

¹University of Louisiana at Lafayette, Lafayette, LA, USA  
²Tampere University, Tampere, Finland  
³University of Jyvaskyla, Jyvaskyla, Finland  
⁴DIGIT, Aarhus University, Aarhus, Denmark  
¹Raju.Gottumukkala@louisiana.edu

ABSTRACT

Affective computing has garnered researchers’ attention and interest in recent years as there is a need for AI systems to better understand and react to human emotions. However, analyzing human emotions, such as mood or stress, is quite complex. While various stress studies use facial expressions and wearables, most existing datasets rely on processing data from a single modality. This paper presents EmpathicSchool, a novel dataset that captures facial expressions and the associated physiological signals, such as heart rate, electrodermal activity, and skin temperature, under different stress levels. The data was collected from 20 participants at different sessions for 26 hours. The data includes nine different signal types, including both computer vision and physiological features that can be used to detect stress. In addition, various experiments were conducted to validate the signal quality.

Introduction

Affective computing is an interdisciplinary field aimed at developing systems that can identify, recognize, and interpret different emotions of human beings. The emotional state of humans is linked to physiological, behavioral, and cognitive changes. These changes depict different emotional traits which can be used to adapt to computer-controlled environments. For example, the current state of a smartphone user can be continuously computed based on their interactions with various digital services. Likewise, certain services can be offered to users based on their current mental state. Stress factors, such as high workload, lack of autonomy, and long working hours, can negatively impact people’s health. In addition, many studies point out that prolonged exposure to stress leads to chronic conditions such as obesity or hypertension, which may exacerbate conditions such as type-II diabetes. Therefore, monitoring and understanding stress in workplaces is essential, especially in professions with increased exposure to stress, often leading to burnout and increased turnover.

Stress is primarily a physiological response to a stimulus, typically triggered by an external factor in an environment. Epinephrine, commonly known as adrenaline, is a significant stress hormone. In a daily routine, a moderate amount of stress is considered beneficial. It causes gentle excitement and improves performance for carrying out regular daily tasks. A reasonable amount of stress that causes excitement is termed eustress or positive stress. On the other hand, undesirable stress is called distress or negative stress. Repeated stress events can lead to emotional exhaustion, and the resulting long-term physiological stress can be harmful; such stress is termed chronic stress. The effects of chronic stress may include anxiety, depression, sleep disturbances, and weight gain. It may even compromise immune responses, resulting in increased vulnerability to several diseases. Studies related to stress have gained interest in recent years due to its widespread effects on health, family, workplace productivity, society, and the economy.

The analysis of stress involves using wearable sensors for deducing physiological signals. Table 1 shows the standard signals provided by wearables. In contrast, the study and analysis of human emotions use a camera for inferring facial expressions representing the emotions. In the literature, emotions and stress signals are typically considered two disjoint topics and are usually studied independently. Schmidt et al. identify the gap in analyzing basic emotions and stress together; however, there is still a need for a dataset that addresses the issue of correlating stress and facial expressions under different stressful situations. This work presents a multimodal stress-emotion dataset containing stress data from wearable devices and emotion data extracted from facial expressions through a video feed. We analyze human expressions and biometric signals under different stress levels.
The dataset presented is the first effort toward identifying stress from non-wearable devices in general, and facial expressions in particular. The current study is inspired by our previous work on stress detection of nurses using wearable devices as well as our initial effort to use facial expressions to study satisfaction. This study collected stress and facial expression data from 20 participants under different stress levels. The dataset is the first of its kind where facial features and physiological features are combined for stress measurement.

| Signal                  | Device                                      | Abbr  |
|-------------------------|---------------------------------------------|-------|
| Heart Activity          | Electrocardiography                          | ECG   |
|                         |                                             | EKG   |
| Skin Response           | Electrodermal Activity                      | EDA   |
| Skin Response           | Galvanic Skin Response                       | GSR   |
| Muscle Activity         | Electromyography                             | EMG   |
| Respiratory Response    | Electromagnetic Generation                  | RR    |
| Respiratory Response    | Respiratory Inductive Plethysmograph        | RIP   |
| Blood Volume Pulse      | Cardiovascular Dynamics                      | BVP   |
|                         |                                             | HR    |
|                         |                                             | HRV   |
| Skin Temperature        | Thermistor                                  | TEMP  |
| Brain Activity          | Electroencephalogram                         | EEG   |
| Eye Activity            | Corneo-retinal Standing                     | EOG   |
| Physical Activity       | 3-axis Accelerometer                        | ACC   |
| Physical Activity       | 3-axis Magnetometer                          | MGM   |
| Heart Activity          | Sphygmomanometer                            | BP    |
| Eye Characteristics     | Pupil Diameter Detector Module              | PD    |
| Body Temperature        | Thermal Imaging                              | TI    |

Table 1. Common Biometric Signals.

1 Background and summary

Affective Computing has recently gathered much interest in the research community. This strong interest is driven by a broad scope of promising technology in many areas, such as virtual reality, intelligent surveillance, and perceptual user interfaces. Affective Computing uses biometric signals to detect the individual’s emotions or state of stress. Stress is one of the big problems in modern society that overloads a significant amount of money on health expenses. Therefore, ambulatory stress detection has become an exciting challenge because early stress detection can mitigate the conversion of stress to a chronic disease and prevent stress-related health issues.

There are a host of surveys in the stress research literature. Sion et al. listed physiological signals that are being used for stress detection. Thapliyal et al. introduced standard stress detection tools and devices. Carneiro et al. described techniques to access and monitor stress in offices and other workplaces. Alberdi et al. provided a comprehensive survey on stress recognition for office environments. They described the signals and their related features as well as successful neural network models that detect stress. Can et al. investigated stress in daily life using wearables and smartphones. They briefly described stress and its impact on society and investigated the stress tests and related signals. However, they did not discuss instruments and devices detecting stress, signal properties, machine learning features, and training parameters. Fukazawa et al. provided a comprehensive survey of stress detection signals and techniques from a machine learning perspective. They only used location, activity, phone usage, context, sleep, and speech features to detect stress. They covered several papers that used a specific signal for stress detection and investigated how commonly a method or signal is used by showing its percentage use across the years. Panicker et al. presented a comprehensive survey on stress detection. They provided a comprehensive description of emotions and their organismic subsystems. They also provided definitions of stress, enumerated stress detection signals and devices, presented three different stages of stress, and evaluated the correlation and differences between different types of stress and emotions. Sun et al. presented an activity-aware stress detection model using Galvanic Skin Response (GSR), Electrocardiogram (ECG), and accelerometer data in three different situations of sitting, standing, and walking for 20 volunteers. They labeled the data based on the task category and achieved 92.4% accuracy for mental stress using decision trees. Zhai and Barreto applied a Support Vector Machine (SVM) to differentiate between stressful and relaxing states.
using Blood Volume Pulse (BVP), Galvanic Skin Response (GSR), Skin Temperature (TEMP), and Pupil Diameter (PD) and achieved 90% accuracy with two-class classification. Kurniawan et al.\textsuperscript{26} investigated several classification methods such as SVM and Gaussian Mixture Models (GMM) to detect stress levels using GSR and speech features. Kurniawan et al.\textsuperscript{26} gained between 70 to 80 percent accuracy using GSR and 92 percent accuracy by combining both the speech and GSR using four classification techniques, namely KMeans, decision tree, Gaussian Mixture Model (GMM) classification models, and Support Vector Machine (SVM). Moreover, some authors use thermal images to detect stress or deception detection\textsuperscript{(G-T)} to do not to drink alcohol or coffee 24 hours before data collection. For participants A-F, no such restrictions were applied. The exclusion criteria were pregnancy, heavy smoking, mental disorders, and chronic or cardiovascular diseases.

The AffectiveRoad dataset\textsuperscript{29} contains biometric signals to study the stress level of 10 drivers while driving in various environments. The study was conducted with drivers taking an 86-minute driving test in Tunisia. Schmidt et al. introduced WESAD dataset\textsuperscript{9} and studied the stress of 15 students while watching a movie and taking a TSST test\textsuperscript{29} using E4 signals. MDPsD (multimodal dataset for psychological stress detection)\textsuperscript{30} is a multimodal stress detection dataset of EDA and PPG signals collected from university students while performing different tests (e.g., TSST\textsuperscript{31}, IQ test and color-word tests\textsuperscript{32}). Mundnich et al.\textsuperscript{33} provided TILES-2018, a multi-sensor dataset that provides a battery of surveys to cover personality traits, behavioral states, job performance, and well-being over time. The SWELL\textsuperscript{34} dataset provides data corresponding to different stress levels of participants while performing some office work (e.g., answering email) at three different stress levels (neutral, stressor, and stressor with interruption) using Kinect, ECG, and emotional expressions extracted from videos. However, this dataset does not provide the videos of the participants and uses outdated facial features to detect emotion and stress. In addition, the SWELL dataset provides the stress level of the entire task as one label.

Our work was inspired by previous work on wearables to monitor physiological signals related to stress. We provide additional modalities to provide a richer dataset containing facial expressions and stress. Our dataset, EmpathicSchool, provides physiological stress signals collected using signal streams from Empatica E4 and facial features from videos of engineering and computer science students performing different tasks. Our primary motivation for creating this dataset was to generate a multimodal emotion-stress detection datum and analyze students’ emotions and stress levels while performing routine tasks like preparing for a presentation or taking an exam. In addition, we analyze the emotion and stress levels of the students after task fulfillment.

2 Methods
This section briefly introduces the sensors used in the data collection process along with data collection protocols. We briefly discuss the methods used to extract features from the video data and provide a summary of sensors used for biomedical data acquisition.

The data was collected in two locations: Tampere University, Finland, and the University of Louisiana at Lafayette, USA. The institutional review board of the University of Louisiana at Lafayette and the Tampere region’s ethics committee approved the study’s protocol for the corresponding locations (SP-21-144-IRI and 50/2021). The data collected at both sites aim to collect facial expressions and biometric signals under different stress levels. The stress levels are determined via National Aeronautics and Space Administration Task Load Index (NASA-TLX) questionnaires\textsuperscript{35, 36}.

2.1 Participants
The participants were recruited via flyers and recruitment emails in the engineering and computer science departments of the University of Louisiana at Lafayette and via email at Tampere University. We asked 35 students to complete the study; however, only 21 students showed up, and after the data collection, one subject asked to be removed from the dataset after completing the data collection. As a result, we collected 20 participants’ data aged from 21 to 35 (M=25.3, STD=4.3). We asked participants (G-T) to do not to drink alcohol or coffee 24 hours before data collection. For participants A-F, no such restrictions were applied. The exclusion criteria were pregnancy, heavy smoking, mental disorders, and chronic or cardiovascular diseases.

2.2 Video data
We asked the participants to perform all the tasks in a sitting position, and the camera captured only their faces and shoulders. The videos of the participants were collected using a 1080p external webcam mounted on top of a laptop screen with 30 frames per second. As approved by the institutional review board at the University of Louisiana at Lafayette, the video data of the participants that consented to release their video files is provided as part of the dataset. In addition to the 1080p resolution video, we also offer a 720p resolution video to make it easier to apply deep learning techniques to the dataset. We post-process the recorded videos to extract facial expressions as discussed in the following subsections. Figure 1 illustrates the experimental setup and the position of the camera.

2.2.1 Facial expression recognition
The facial expression recognition system constitutes two modules. In the first module, we use the Haar cascade\textsuperscript{37} frontal face detection algorithm implemented in the OpenCV\textsuperscript{38} library to detect the face in the video frame. Haar cascade uses Haar-like
features to encode the local appearance of faces\textsuperscript{39}. The detected face is processed and passed through a pre-trained model to recognize facial expressions in the second module. We used MiniXception\textsuperscript{40}, trained over Facial Expression Recognition 2013 (FER-2013) dataset\textsuperscript{41} for facial expression recognition. FER-2013 dataset contains 28,709 training images, 3,589 validation images, and 3,589 test images from the seven basic facial expression categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). Our experiments deduced the facial expressions from all the video recordings. This data is provided them as part of the dataset.

2.2.2 Dlib features
Facial landmarks are used in various computer vision tasks, such as drowsiness detection\textsuperscript{42}, fatigue detection\textsuperscript{43}, facial expression recognition\textsuperscript{44}, and micro-expression classification\textsuperscript{45}. Therefore, we provide the facial landmarks of all the video frames collected as part of the dataset. We utilize Dlib\textsuperscript{46}, which returns the position of 68 facial landmarks. An example of landmark detection on a facial image based on Dlib is shown in Figure 2.

2.3 Biometric data
We used the E4 wristband (Empatica Inc., Milano, Italy) for physiological data acquisition. E4 is a medical-grade wearable device that offers real-time acquisition of EDA, heart rate, skin temperature, and accelerometer data from the subject’s right hand. EDA is measured via E4’s silver (Ag) electrode (valid range [0.01–100] µS), while heart rate is measured via E4’s photoplethysmographic (PPG) sensor. The E4 wristband is powered by a rechargeable lithium battery and transmits data to the subject’s smartphone using Bluetooth. All the data collected from the E4 wristband and the sampling frequencies are presented in Table 2.

The biometric signals that we measured via the E4 wristband and are provided in our dataset are the following:

- **Heart rate:** The heart rate of a healthy individual, irrespective of gender, ranges from 60 to 100 beats per minute in a resting state. However, the heart rate varies significantly with activity or emotional state\textsuperscript{47}. Heart rate is generally associated with stressful situations, but a high heart rate should not be interpreted as high stress.

- **Skin temperature:** Skin temperature varies for various activities due to skin blood flow. Skin temperature ranges from 33.5°C to 36.93°C\textsuperscript{48}. However, this can vary quite widely based on the type and length of activity and indoor room temperature.
### Table 2. Signals and the sampling frequency of Empatica E4 sensors

| Signal                     | Abbreviation | Frequency |
|----------------------------|--------------|-----------|
| Electrodermal Activity     | EDA          | 4.0 Hz    |
| Heart Rate                 | HR           | 1.0 Hz    |
| Skin Temperature           | ST           | 1.0 Hz    |
| Accelerometer              | ACC          | 32 Hz     |
| Inter-Beat Interval        | IBI          | ---       |
| Blood Volume Pulse         | BVP          | 64 Hz     |

- **Electrodermal activity:** EDA measures the amount of sweat glands’ activity by calculating the skin’s electrical conductance using silver-chloride electrodes. The EDA signal is measured in units of micro-siemens (µS). Stadler et al.\(^49\) mention that EDA peaks are event-related and can be a good estimator of the body’s response to the stimulus.

- **Accelerometer data:** Accelerometer sensors can be used for multiple tasks (e.g., human action recognition and step counting). By measuring orientation and acceleration force, the accelerometer sensor can determine the device’s orientation (horizontal or vertical) and the type of movement. The accelerometers vary in type (digital and analog), sensitivities, and the number of axes. Our dataset provides three-axis accelerometer data that can measure the orientation of the sensor in three dimensions which is able to capture the activity more precisely.

- **Blood Volume Pulse:** Blood volume pulse is measured by using a photoplethysmograph (PPG) that measures light absorption to find the volume peaks, and heartbeats. The heart rate is measured based on the time intervals from two consequent beats. Moreover, the heart rate variability can be measured by BVP signal.

- **Inter Beat Interval:** The inter-beat interval is the time interval between individual heartbeats. IBI is the time difference between the two consequent peaks in seconds that differs from beat to beat. Heart rate variability is derived from the IBI signal.

### 2.4 Data collection protocol

Our stress detection study was conducted in a laboratory setting. We aimed to detect students’ stress during their typical duties in a laboratory setting while sitting behind a computer wearing a wristband. We collected data from 20 participants under different stress levels (normal and high stress). The normal stress level was assumed when the participants were at rest or performing a task that did not require significant mental effort. For the stressful tasks, we asked participants to accomplish the given assignment in a limited period of time\(^34\). Each of these tasks are supposed to be performed for 10 minutes. However, due to data synchronization issues, 30 seconds of data is trimmed from the beginning and ending of each task. This gives us 9 minutes worth of information for each participant and each task. The actions for nine different data collection sessions (tasks) and the duration of each session are provided in Table 3. All the subjects used the complete allotted time for all the tasks, except, eight out of the twenty participants decided to complete Delivering the Presentation Task (T3) earlier than the allotted time. In these cases, we stored only the data of the task duration. In addition, Table 3 gives the expected stress level for each task.

| Task | Action performed               | Duration (min) | Expected Stress Level |
|------|--------------------------------|----------------|-----------------------|
| T1   | Reading a magazine             | 9              | Normal                |
| T2   | Preparing a presentation       | 9              | Stressed              |
| T3   | Delivering the presentation    | 5-9            | Stressed              |
| T4   | Rest and recovery              | 9              | Normal                |
| T5   | IQ test / Stroop Color-Word Test | 9          | Stressed              |
| T6   | Listening to calm music        | 9              | Normal                |
| T7   | Watching amusing video(s)      | 9              | Amused                |
| T8   | Controlled breathing exercise  | 9              | Normal                |
| T9   | Recovery                       | 9              | Normal                |

*Table 3. List of tasks performed by participants, the total duration of the data included in the release, and the expected stress level for each task.*

In task T1, the students in the United States were asked to find and read an arbitrary article from Ars Technica or Wired Internet magazines, while in the Finland site, the students were asked to read an arbitrary magazine from the Internet according
to their choice to reduce their stress level to achieve a baseline level of stress\textsuperscript{50}. In task T2, the participants were asked to prepare for a presentation in a limited time, followed by presenting task T3 in the allocated time. We considered both tasks (preparing and delivering presentations) as stressful tasks. In the next task (T4), participants were asked to take a rest. In task T5, the Stroop Color-Word or IQ test was taken by each participant to observe their stress level while taking tests in a limited time period. In task T6, we asked the participants to listen to calm music, and we considered them to not to be stressed during this activity\textsuperscript{51}. The rest of the data collection included watching an amusing video (T7) followed by one or two different rest activities, namely controlled breathing exercises (T8) and a recovery period (T9). Six out of twenty participants were asked to have a recovery task after a controlled breathing exercise. The video and wristband data have different start and end times, so we trimmed 60 seconds of each task for data length consistency after time synchronization (30 seconds from the beginning and the end). Participants (G-T) did not attempt the recovery task (T9).

An Empatica E4 was worn on the wrist of the dominant arm, and a camera collected facial videos during the studies. During the sessions, we continuously collected the students’ physiological data and facial features. We detected stress events while monitoring the biometric signal streams of the student. Moreover, for participants G-T, we asked the student to validate their stress level during each stressed task by completing the NASA questionnaire. The participants were also asked to complete a questionnaire if they experienced stress during a rest (normal stress) task.

Because of COVID-19 restrictions, participants were completely alone during the tests, the test and presentations were presented virtually, and the recruiter was in another room during the tests. Figure 3 presents the framework for the data collection. A detailed explanation of each of the collected data is given below.

2.5 Stress detection surveys
In this study, we provided NASA-TLX forms to participants G-T and asked them to complete the questionnaires after each stressful task. Each task was divided into 5 two-minute sub-intervals except for the first and last sub-intervals span over 90 seconds (1.5 minutes) adding up to the 9 minutes time for each task. Therefore, Participants (G-T) were asked to determine their overall stress level at each sub-interval during stressful tasks. In addition, we also asked participants (G-T) to rate their overall workload during different stressful tasks and provided that information and labels in the dataset. Table 4 shows the NASA questionnaires that were asked to evaluate a participant’s stress levels from 0 to 20 for each sub-interval.

3 Data records
The dataset contains 20 folders (A to T) with each participant’s data separately. The dataset is anonymized and will be made available upon request to the corresponding author. Each folder consists of nine sub-folders (T1 to T9) according to the tasks performed by participants. Table 5 shows the data availability for the participants in each folder. In addition, each folder
| No | Category         | Question                                                                 |
|----|------------------|--------------------------------------------------------------------------|
| 1  | Mental Demand    | How mentally demanding was the task?                                     |
| 2  | Physical Demand  | How physically demanding was the task?                                   |
| 3  | Temporal Demand  | How hurried or rushed was the pace of the task?                          |
| 4  | Performance      | How successful were you in accomplishing what you were asked to do?      |
| 6  | Effort           | How hard did you have to work to accomplish your level of performance?   |
| 7  | Frustration      | How insecure, discouraged, irritated, stressed, and annoyed were you?    |

Table 4. Workload questionnaire to evaluate the overall stress level of students during different tasks.

| ID/Task | T1 | T2 | L2 | T3 | L3 | T4 | T5 | T6 | T7 | L7 | T8 | T9 | Video | labels | Site |
|---------|----|----|----|----|----|----|----|----|----|----|----|----|-------|--------|------|
| A       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | Finland |
| B       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | Finland |
| C       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | Finland |
| D       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | Finland |
| E       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | Finland |
| F       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | Finland |
| G       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | United States |
| H       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | United States |
| I       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | United States |
| J       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | United States |
| K       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | United States |
| L       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | United States |
| M       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | United States |
| N       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | United States |
| O       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | United States |
| P       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | United States |
| Q       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | United States |
| R       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | United States |
| S       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | United States |
| T       | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓  | ✓     |        | United States |

Table 5. The availability of the tasks for each of the participants (✓ mark corresponds to the case where the data is not available for the participant, ✓ shows the availability of the data).

follows these naming conventions: The first letter of the filename shows the subject ID, followed by two letters showing the task number (e.g., T1). For example, LX corresponds to the validated stress labels of the task TX.

We survey students for every 2 minutes intervals to measure their stress during stressed tasks. However, it was impossible to ask the participants to complete the NASA forms for each interval (extreme survey bias due to predictability of the questions), so we asked the participants to complete the NASA-TLX forms to measure the overall workload of the stressed task. However, we show LX as the availability of the user’s stress level for the two-minute intervals while the labels survey the participants’ overall workload of each task. The labels (overall and intervals) are provided in the stress_level.csv files. The NASA-TLX survey asked participants to provide overall feedback about the entire task on physical demand, mental demand, temporal demand, frustration, effort, and performance.

3.1 Data files description
The following is the description of files in the task sub-folders:

- **Xception.xlsx**: The Excel file contains the facial expressions deduced at frame level for each task using MiniXception, trained over the FER-2013 dataset. The details are given in section 2.2.1.

- **HR.csv**: The Excel file contains two columns, including the HR signal of the participant and the corresponding data collection time in seconds, respectively.

- **EDA.csv**: The Excel file contains the EDA signal collected from the wristband and the corresponding data collection time in seconds.
• **TEMP.csv**: The Excel file contains the participants’ skin temperature (TEMP) collected from the wristband along with the corresponding data collection time in seconds.

• **Acc.csv**: The Excel file contains four columns corresponding to the x-axis, y-axis, and z-axis accelerometer data, respectively, along with the collection time in seconds.

• **V.mp4**: The video file contains raw video files of participants that consented to release their videos. This file is not available for all participants.

• **Stress_level.csv**: The Excel file holds the validated stress level of the participants extracted from their answers. Each label is for a two-minute interval (represented 0, 1, 2, and unknown where 0 = no stress; 1=low stress; 2=high stress, unknown for unknown stress level).

• **Landmarks.csv**: The Excel file contains the normalized position of the participant’s 68 facial landmarks (Dlib features described in Section 2.2.2) over time.

• **BVP.csv**: The Excel file contains two columns corresponding to the BVP signal and the time of data collection in seconds.

• **IBI.csv**: The Excel file contains two columns of IBI data; the data is directly collected from the E4 device and is available in the first task folder; however, it contains data of the entire study.

We provide the time stamps for the biometric signals as a separate column to facilitate the synchronization of the physiological signals and video. E4 provides BVP and IBI signals that we did not use in the models, and the data is available in the raw data zip files. The videos and biometric signals have exact start times, and can be synced by interpolation, downsampling, or upsampling data processing methods.

4 Technical validation

4.1 Metrics
In this paper we compare the performance of classification methods using below metrics:

• **Precision** represents the closeness of the predicted results:

\[
Precision = \frac{TP}{TP + FP}.
\]  

4.2 Machine learning and stress detection model
The four signals and video are collected at different sampling rates due to the variation of sensors. The frequency of the signals ranges from 1 to 30 HZ for heart rate and video data, respectively. We used the biometric signals and facial expressions of video as features for our stress detection model. We decided to use a frequency of 4 HZ after evaluating the accuracy of different models to minimize the information loss while monitoring the computational cost of the models. Moreover, models with repeated data samples are prone to overfitting due to the increased chance of spotted test samples by the model during the training. The code associated with the data preprocessing, upsampling, and down sampling is available on GitHub repository. The machine learning classification models process the biometric signals and facial expression features one-time step at a time. The stress detection models are trained and tested based on the participants’ survey answers. The output of the stress detection models is the stress level of the participant at the mentioned time step (0: No stress, 1: medium stress, 2: high stress).
4.3 Physiological signals vs. facial expressions

We analyzed the physiological signals obtained from the Empatica E4 watch and the predicted facial expressions from the video data for different tasks. Since the frame rate for the video is higher than the frequency of the physiological signals, we interpolated the physiological data to make it equal in length to the length of deduced emotions from the video frames. To smoothen the emotion data (Fig 4), we use a quadratic polynomial used in the Savitzky-Golay method\textsuperscript{54}. The smoothing filter was applied to data at the task level by keeping the span of the smoothing filter at 30 percent. After applying the smoothing filter, we scaled the data between 0 and 1 by applying $\tilde{x}_i = \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}$, where $\tilde{x}_i$ is the scaled value, $x_{\text{min}}$ is the minimum, and $x_{\text{max}}$ is the maximum smoothed value at the corresponding task/session for the participant. Fig 4 shows the happiness curves and physiological signals of subjects A-F performing different tasks (T1-T9). The happiness curve is observed to fluctuate more in T5 and T7 for subjects A-F as compared to other tasks. For a given task, the temperature is observed to stay steady within the task; however, some difference has been observed in the temperature values between the tasks. EDA curve is observed to stay steady during T1 and T2, while slight variation is observed in T3. We also observed a peak heart rate value for subjects B, C, and D for T3. These observations gave a brief insight into the data and validated the assumption that different tasks have different natural influences on the expressions and biometric signals; nevertheless, they alone cannot explain the reasons for changes or any correlation in the curves. Similar plots can be generated for comparing other emotions (Angry, Disgust, Fear, Sad, Surprise) to the physiological signals provided in the dataset.

Figure 5 shows the overall stress level of the students performing different tasks based on participants’ feedback. For example, the overall stress level of delivering a presentation was higher compared to the other tasks. In addition, the stress level of preparing for a presentation showed similar results when taking a test for the participants.

A stress detection algorithm is designed to detect stress from facial expressions and biometric signals. In order to extract the features for the stress detection model, we first perform facial expression recognition in the video, this process is described in section 2.2.1. Then, a sliding window of 20 seconds and a step-size of ten seconds is used to extract statistical features from the data. In total we used 28 features extracted from both video and biometric signals. The mean of the facial expressions extracted from video is concatenated with the features extracted from bio-metric signals. For each of the 3 biometric signals, we calculate the mean, minimum, maximum, and standard deviation along the 20 second sliding window. In addition, we extract the kurtosis, skewness, number of peaks, amplitude and duration of the peak for the EDA signal. For HR, we extract the root mean square (quadratic mean), number of peaks, amplitude, and duration of the peaks. These statistical features for biometric signals are extracted based on an earlier work by Neska et al\textsuperscript{35}. The labels for each window are calculated based on the average stress values during the time window: ‘no stress’ if $S \leq 0.65$, ‘medium stress’ if $0.65 < S \leq 1.3$, and ‘high stress’ if
$S > 1.3$. The source code for the stress detection algorithm comprises of feature extraction, stress detection, and change-point detection is provided in the GitHub repository.

A machine learning model is trained from the features extracted in the second step to predict stress during the twenty-second window. To evaluate the effectiveness of the model and avoid overfitting, we followed a leave one subject out strategy on the test dataset. Where the model is trained on data from 11 participants and evaluated against the 12th. The ground truth for the participants is available for 12 participants. This process is repeated for all the 12 participants. In the data, the stress labels are imbalanced (low stress: 76%, medium stress: 16%, high stress: 8%). We employed a KNN oversampling approach to balance the data. We tested three different machine learning algorithms: random forests, decision trees, and XGBoost to classify the stress values of the participants for different sets of features. The precision, recall, and f-score of these three approaches are presented in Table 6.

The emotions model uses the mean of the facial expressions extracted from video to detect stress. The biometrics model uses the 21 features to train the machine learning model. The multimodal analysis uses 28 features to predict the stress level in our data. Larger number of features leads to a bad model that is overfits and is not generalizable for unseen data. To reduce the number of features, we use Pearson correlation analysis to select the top 10 features correlated with stress. The top 10 features we observed are: Disgust, Scared, Surprised from facial expressions and the kurtosis of EDA, skewness of EDA, HR standard deviation, HR root mean squared (geometric mean), HR number of peaks, hr amplitude, and hr duration. We train a machine learning model with these 10 features to predict the value of stress. Our results indicate that the reduced set of features improved the performance of stress prediction by 5.9% for random forest and 3.7% for decision trees, but led to a reducing by 1.2% for XGBoost for recall. Similar observations can be made for precision and f1_score as well. However, the multimodal (top ten) features, i.e., combining both emotion and biometric features improves the model compared to using just emotion or biometric data. In addition, while the machine learning models are able to generalize for detecting stress, we also observed that the recall is lower for certain individuals, where the machine learning models predict low stress value all the time. The code to train the machine learning models and evaluate the results are available on GitHub.

We computed the relationship between all the variables to stress using a one-way ANOVA sample test. First, we normalized the signals over each subject separately to reduce the effect of variation in biometric signals’ amplitude among subjects. Then we divided stress labels from 1 to 20 to three different classes (0: 0 to 6, 1: 7 to 13, 2: 14 to 20) to reduce the degrees of freedom. Finally, for each stress level, we averaged the signals for all the tuples and computed the ANOVA score. The null hypothesis is that there is no difference between the signal and the stress levels. The results of the ANOVA test are presented in Table 7. With the ANOVA statistic of 4.35 and the p-value of 0.03, we can reject the null hypothesis that there is no difference in heart rate with respect to stress.

Emotions like angry, surprise, and sad are highly correlated with stress. This relationship between heart rate, emotions, and stress have been previously discussed by several authors. Earlier studies on stress detection evaluated the relationship
between stress and skin temperature\textsuperscript{61,62}. Skin temperature drops during the onset of stress and increases above the normal temperature after the onset\textsuperscript{25,63}. Therefore, skin temperature and stress are not directly correlated. However, a sudden decrease in skin temperature is a good indicator of early stress, but during later time periods, the skin temperature is higher. The machine learning models can use these complex relationships to identify high-stress levels, as represented in Table 6.

| Model                  | Features | Recall                | Precision               | F1_score               |
|------------------------|----------|-----------------------|-------------------------|------------------------|
| **Random Forest**      |          |                       |                         |                        |
| Emotions               | 7        | 62.53±10.28           | 77.51±12.74             | 66.65±13.57            |
| Biometric              | 21       | 57.41±25.74           | 72.81±29.61             | 59.57±29.42            |
| Multimodal (All)       | 28       | 63.24±21.29           | 78.55±20.50             | 66.54±24.50            |
| Multimodal (Top 10)    | 10       | 69.35±12.73           | 85.04±10.04             | 75.93±10.57            |
| **Decision Tree**      |          |                       |                         |                        |
| Emotions               | 7        | 52.68±14.48           | 77.51±12.74             | 66.65±24.50            |
| Biometric              | 21       | 43.20±25.17           | 59.75±30.30             | 42.64±29.80            |
| Multimodal (All)       | 28       | 49.72±19.95           | 51.56±31.56             | 48.25±26.74            |
| Top 10                 | 10       | 63.40±12.62           | 73.70±14.26             | 66.08±12.30            |
| **XGBoost**            |          |                       |                         |                        |
| Emotions               | 7        | 61.75±8.40            | 69.68±14.57             | 63.74±10.66            |
| Biometric              | 21       | 64.51±17.77           | 72.88±23.88             | 67.69±20.31            |
| Multimodal (All)       | 28       | 65.91±18.30           | 79.41±21.69             | 70.42±19.88            |
| Multimodal (Top 10)    | 10       | 63.14±10.38           | 69.53±7.05              | 65.61±7.56             |

Table 6. Performance of different models using balanced data.

| No. | Signal   | Significant | F   | p-value |
|-----|----------|-------------|-----|---------|
| 1   | Angry    | True        | 8.481 | 0.004   |
| 2   | Surprised| True        | 5.270 | 0.022   |
| 3   | Sad      | True        | 4.723 | 0.030   |
| 4   | HR       | True        | 4.346 | 0.037   |
| 5   | EDA      | False       | 1.531 | 0.216   |
| 6   | Happy    | False       | 1.365 | 0.243   |
| 7   | Neutral  | False       | 0.082 | 0.775   |
| 8   | Scared   | False       | 0.032 | 0.859   |
| 9   | Disgust  | False       | 0.029 | 0.865   |
| 10  | TEMP     | False       | 0.001 | 0.981   |

Table 7. One-way ANOVA significance test results regarding stress.

5 Limitations and future works

We collected data in laboratory conditions, and the below items are the limitations of our stress detection dataset:

- Not all of the dataset is covered by stress labels. Unlabelled data does not necessarily imply a lack of stress; it just means that we did not decide to collect the labels in assumed low-stress tasks. However, we asked participants to let us know if they experienced the onset of stress in low-stress tasks. Except for one participant, every subject did not experience stressful moments in the low-stress tasks.

- The stress labels have granularity of 2-minutes. Finer, granularity can lead to recall bias\textsuperscript{64}.

- This dataset is primarily focused on labeling acute stress among participants and does not consider chronic stress as a factor in the participants. There is not sufficient research on chronic stress detection using biometric signals.

- We conducted a laboratory setting for stress detection and the participants may experience some stress due to the new environment. However, we asked the participants to read magazines under task 1 to mitigate the newcomer’s socialization through the Lens of Stress effect\textsuperscript{65}.
6 Usage notes

The EmpathicSchool dataset can be used for various applications. We envision the following potential applications and further analyses for the following research communities.

Stress management: The EmpathicSchool dataset can be used to analyze human facial expressions in co-relation with stress. The stress management applications help improve humans’ everyday functioning and mitigate chronic stress.

Emphatic building: EmpathicSchool dataset consists of data at different stress levels (sessions), which are analogous to different events in daily life in the working space. The dataset is a stepping stone toward analyzing employees’ well-being in an Emphatic building environment.

Perceptual interface: The dataset is also useful for further studies in developing perceptual interface for patients in Intensive Care Unit (ICU). The analysis of human expressions in co-relation with physiological signals can be useful for pain management and similar protocols can be used for.

Online education: In online education, it is important to adapt the style and comfort according to the students needs. To render students satisfaction, frustration or stress levels, the EmpathicSchool dataset can be used to develop such applications.

7 Code availability

We provide the code for deducing facial expressions from video data using InceptionV3\(^\text{66}\) model pre-trained on FER-2013 dataset\(^\text{41}\), and dlib features\(^\text{46}\) that provid 68 facial landmarks on a facial image as discussed in section 2.2.2. We also provide codes for analysing (producing the plots/curves) physiological signals vs. facial expressions. The code is publicly available on GitHub\(^\text{53}\) and the data will be available upon request to the corresponding author.

Acknowledgements

This work has been supported by NSF IUCRC CVDI, project AMALIA funded by Business Finland and DSB, as well as projects Mad@work and Stroke-Data funded by Haltian.

References

1. Poria, S., Cambria, E., Bajpai, R. & Hussain, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion* **37**, 98–125 (2017).
2. Politou, E., Alepis, E. & Patsakis, C. A survey on mobile affective computing. *Comput. Sci. Rev.* **25**, 79–100 (2017).
3. Peternel, K., Pogačnik, M., Tavčar, R. & Kos, A. A presence-based context-aware chronic stress recognition system. *Sensors* **12**, 15888–15906 (2012).
4. Bickford, M. Stress in the workplace: A general overview of the causes, the effects, and the solutions. *Can. Mental Heal. Assoc. Nfld. Labrador Div.* **8**, 1–3 (2005).
5. Wellen, K. E., Hotamisligil, G. S. et al. Inflammation, stress, and diabetes. *The J. clinical investigation* **115**, 1111–1119 (2005).
6. Greenglass, E. R., Burke, R. J. & Fiksenbaum, L. Workload and burnout in nurses. *J. community & applied social psychology* **11**, 211–215 (2001).
7. Li, C.-T., Cao, J. & Li, T. M. Eustress or distress: An empirical study of perceived stress in everyday college life. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, 1209–1217 (2016).
8. Maslach, C. A multidimensional theory of burnout. *Theor. organizational stress* **68**, 16 (1998).
9. Schmidt, P., Reiss, A., Duerichen, R., Marberger, C. & Van Laerhoven, K. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, 400–408 (2018).
10. Hosseini, S. et al. A multi-modal sensor dataset for continuous stress detection of nurses in a hospital. *arXiv preprint arXiv:2108.07689* (2021).
11. Sohrab, F., Raitoharju, J. & Gabbouj, M. Facial expression based satisfaction index for empathic buildings. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, 704–707 (2020).

12. Picard, R. W., Wexelblat, A. & Clifford I. Nass, C. I. N. I. Future interfaces: social and emotional. In *CHI’02 Extended Abstracts on Human Factors in Computing Systems*, 698–699 (2002).

13. Tao, J. & Tan, T. Affective computing: A review. In Tao, J., Tan, T. & Picard, R. W. (eds.) *Affective Computing and Intelligent Interaction*, 981–995 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2005).

14. Can, Y. S., Arnrich, B. & Ersoy, C. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *J. biomedical informatics* **92**, 103139 (2019).

15. Plarre, K. *et al.* Continuous inference of psychological stress from sensory measurements collected in the natural environment. In *Proceedings of the 10th ACM/IEEE international conference on information processing in sensor networks*, 97–108 (IEEE, 2011).

16. Smets, E. *et al.* Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *NPJ digital medicine* **1**, 1–10 (2018).

17. Wijsman, J., Grundlehner, B., Liu, H., Hermens, H. & Penders, J. Towards mental stress detection using wearable physiological sensors. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1798–1801 (IEEE, 2011).

18. Sioni, R. & Chittaro, L. Stress detection using physiological sensors. *Computer* **48**, 26–33 (2015).

19. Thapliyal, H., Khalus, V. & Labrador, C. Stress detection and management: A survey of wearable smart health devices. *IEEE Consumer Electron. Mag.* **6**, 64–69 (2017).

20. Carneiro, D., Novais, P., Augusto, J. C. & Payne, N. New methods for stress assessment and monitoring at the workplace. *IEEE Transactions on Affect. Comput.* **10**, 237–254 (2017).

21. Alberdi, A., Azpiria, A. & Basarab, A. Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. *J. biomedical informatics* **59**, 49–75 (2016).

22. Fukazawa, Y. *et al.* Smartphone-based mental state estimation: A survey from a machine learning perspective. *J. Inf. Process.* **28**, 16–30 (2020).

23. Panicker, S. S. & Gayathri, P. A survey of machine learning techniques in physiology based mental stress detection systems. *Biocybern. Biomed. Eng.* **39**, 444–469 (2019).

24. Sun, F.-T. *et al.* Activity-aware mental stress detection using physiological sensors. In *International conference on Mobile computing, applications, and services*, 282–301 (Springer, 2010).

25. Zhai, J. & Barreto, A. Stress detection in computer users based on digital signal processing of noninvasive physiological variables. In *2006 international conference of the IEEE engineering in medicine and biology society*, 1355–1358 (IEEE, 2006).

26. Kurniawan, H., Maslov, A. V. & Pechenizkiy, M. Stress detection from speech and galvanic skin response signals. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, 209–214 (IEEE, 2013).

27. Pavlidis, I. System and method using thermal image analysis for polygraph testing (2005). US Patent 6,854,879.

28. Haouij, N. E., Poggi, J.-M., Sevestre-Ghalila, S., Ghozi, R. & Jaidane, M. Affectiveroad system and database to assess driver’s attention. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, 800–803 (2018).

29. Kirschbaum, C., Pirke, K.-M. & Hellhammer, D. H. The ‘trier social stress test’–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology* **28**, 76–81 (1993).

30. Chen, W., Zheng, S. & Sun, X. Introducing mdpsd, a multimodal dataset for psychological stress detection. In *Big Data: 8th CCF Conference, BigData 2020, Chongqing, China, October 22–24, 2020, Revised Selected Papers*, vol. 1320, 59 (Springer Nature, 2021).

31. Birkett, M. A. The trier social stress test protocol for inducing psychological stress. *JoVE (Journal Vis. Exp.*) e3238 (2011).

32. Scarpina, F. & Tagini, S. The stroop color and word test. *Front. psychology* **8**, 557 (2017).

33. Mundnich, K. *et al.* Tiles-2018, a longitudinal physiologic and behavioral data set of hospital workers. *Sci. Data* **7**, 1–26 (2020).
34. Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M. A. & Kraaij, W. The swell knowledge work dataset for stress and user modeling research. In Proceedings of the 16th international conference on multimodal interaction, 291–298 (2014).
35. Hart, S. G. Nasa-task load index (nasa-tlx): 20 years later. In Proceedings of the human factors and ergonomics society annual meeting, vol. 50, 904–908 (Sage publications Sage CA: Los Angeles, CA, 2006).
36. Hart, S. G. & Staveland, L. E. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Advances in psychology, vol. 52, 139–183 (Elsevier, 1988).
37. Viola, P. & Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, vol. 1, I–I (Ieee, 2001).
38. Bradski, G. The opencv library. Dr. Dobb’s Journal: Softw. Tools for Prof. Program. 25, 120–123 (2000).
39. Padilla, R., Costa Filho, C. & Costa, M. Evaluation of haar cascade classifiers designed for face detection. World Acad. Sci. Eng. Technol. 64, 362–365 (2012).
40. Arriaga, O., Valdenegro-Toro, M. & Plöger, P. Real-time convolutional neural networks for emotion and gender classification. arXiv preprint arXiv:1710.07557 (2017).
41. Goodfellow, I. J. et al. Challenges in representation learning: A report on three machine learning contests. In International conference on neural information processing, 117–124 (Springer, 2013).
42. Mohanty, S., Hegde, S. V., Prasad, S. & Manikandan, J. Design of real-time drowsiness detection system using dlib. In 2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), 1–4 (IEEE, 2019).
43. Irtija, N., Sami, M. & Ahad, M. A. R. Fatigue detection using facial landmarks. In International Symposium on Affective Science and Engineering ISASE2018, 1–6 (Japan Society of Kansei Engineering, 2018).
44. Heidari, N. & Iosifidis, A. Progressive spatio-temporal bilinear network with monte carlo dropout for landmark-based facial expression recognition with uncertainty estimation. In IEEE International Workshop on Multimedia Signal Processing (IEEE, 2021).
45. Kumar, A. J. R. & Bhanu, B. Micro-expression classification based on landmark relations with graph attention convolutional network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021).
46. King, D. E. Dlib-ml: A machine learning toolkit. The J. Mach. Learn. Res. 10, 1755–1758 (2009).
47. Shi, H. et al. Differences of heart rate variability between happiness and sadness emotion states: a pilot study. J. Med. Biol. Eng. 37, 527–539 (2017).
48. Bierman, W. The temperature of the skin surface. J. Am. Med. Assoc. 106, 1158–1162 (1936).
49. Stadler, R., Jepson, A. S. & Wood, E. H. Electrodermal activity measurement within a qualitative methodology: Exploring emotion in leisure experiences. Int. J. Contemp. Hosp. Manag. (2018).
50. Jin, P. Efficacy of tai chi, brisk walking, meditation, and reading in reducing mental and emotional stress. J. psychosomatic research 36, 361–370 (1992).
51. Ferrer, E. et al. Playing music to relieve stress in a college classroom environment. Coll. Stud. J. 48, 481–494 (2014).
52. Hellevik, O. Extreme nonresponse and response bias. Qual. & Quant. 50, 1969–1991 (2016).
53. CPHSLab. Empathicschool. https://github.com/CPHSLab/EmpathicSchool (2022).
54. Schafer, R. W. What is a savitzky-golay filter?[lecture notes]. IEEE Signal processing magazine 28, 111–117 (2011).
55. El Haouij, N., Poggi, J.-M., Ghozi, R., Sevestre-Ghalila, S. & Jaïdane, M. Random forest-based approach for physiological functional variable selection for driver’s stress level classification. Stat. Methods & Appl. 28, 157–185 (2019).
56. Kim, T. K. Understanding one-way anova using conceptual figures. Korean journal anesthesiology 70, 22 (2017).
57. Taelman, J., Vandeput, S., Spaepen, A. & Huffel, S. V. Influence of mental stress on heart rate and heart rate variability. In 4th European conference of the international federation for medical and biological engineering, 1366–1369 (Springer, 2009).
58. Vrijkotte, T. G., Van Doornen, L. J. & De Geus, E. J. Effects of work stress on ambulatory blood pressure, heart rate, and heart rate variability. Hypertension 35, 880–886 (2000).
59. Schubert, C. et al. Effects of stress on heart rate complexity—a comparison between short-term and chronic stress. Biol. psychology 80, 325–332 (2009).
60. Goodie, J. L., Larkin, K. T. & Schauss, S. Validation of polar heart rate monitor for assessing heart rate during physical and mental stress. *J. Psychophysiol.* **14**, 159 (2000).

61. Chen, J., Abbod, M. & Shieh, J.-S. Pain and stress detection using wearable sensors and devices—a review. *Sensors* **21**, 1030 (2021).

62. Li, R. & Liu, Z. Stress detection using deep neural networks. *BMC Med. Informatics Decis. Mak.* **20**, 1–10 (2020).

63. Yamakoshi, T. *et al.* Feasibility study on driver’s stress detection from differential skin temperature measurement. In *2008 30th annual international conference of the ieee engineering in medicine and biology society*, 1076–1079 (IEEE, 2008).

64. Tarrant, M. A., Manfredo, M. J., Bayley, P. B. & Hess, R. Effects of recall bias and nonresponse bias on self-report estimates of angling participation. *North Am. J. Fish. Manag.* **13**, 217–222 (1993).

65. Allen, T. D., McManus, S. E. & Russell, J. E. Newcomer socialization and stress: Formal peer relationships as a source of support. *J. Vocat. Behav.* **54**, 453–470 (1999).

66. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826 (2016).