Computers in Biology and Medicine 186 (2025) 109555

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Stress classification with in-ear heartbeat sounds

Danielle Benesch [a,d], Bérangère Villatte [b], Alain Vinet [c], Sylvie Hébert [b], Jérémie Voix [a,d], Rachel E. Bouserhal [a,d,*]

[a] École de technologie supérieure, 1100 Notre-Dame St W, Montreal, H3C 1K3, Quebec, Canada
[b] School of Speech-Language Pathology and Audiology, Université de Montréal, 7077 Av du Parc, Montreal, H3N 1X7, Quebec, Canada
[c] Centre de recherche en physiologie cardiovasculaire, Hopital du Sacré-Coeur-de-Montréal, 5400, boul. Gouin Ouest, Montreal, H4J 1C5, Quebec, Canada
[d] Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT), 527 Rue Sherbrooke O #8, Montréal, QC H3A 1E3, Canada

ARTICLE INFO

Dataset link: https://github.com/critias-ca/in-ear-stress

Keywords:
Stress monitoring
Heart rate variability
In-ear audio
Hearables
Tinnitus
Hyperacusis
Misophonia

ABSTRACT

Background: Although stress plays a key role in tinnitus and decreased sound tolerance, conventional hearing devices used to manage these conditions are not currently capable of monitoring the wearer's stress level. The aim of this study was to assess the feasibility of stress monitoring with an in-ear device.
Method: In-ear heartbeat sounds and clinical-grade electrocardiography (ECG) signals were simultaneously recorded while 30 healthy young adults underwent a stress protocol. Heart rate variability features were extracted from both signals to train classification algorithms to predict stress vs. rest.
Results: Models trained and tested using in-ear heartbeat sounds appeared to perform better than the models trained and tested using the ECG signals. However, further analyses comparing heart rate variability features extracted from ECG and the in-ear heartbeat sounds suggest that the improvement in stress prediction performance was driven by the increased presence of artifacts (e.g. movement or speech) during the stress tasks, rather than physiologically meaningful changes in the heartbeat signals that would be indicative of stress in real-world settings. To address this difference in error between rest and stress conditions, a data augmentation method was proposed to balance the error.
Conclusions: The final system demonstrates the viability of robust stress recognition with only in-ear heartbeat sounds, which could expand the capabilities of hearing devices used to address conditions related to stress and noise. The proposed data augmentation method effectively identified and addressed artifact-related biases, which could broadly be applied to improve robustness of biosignal monitoring with machine learning.

1. Introduction

Psychological stress is closely associated with tinnitus, the subjective perception of sound in the absence of an external source [1], and decreased sound tolerance, the intolerance to sounds that do not bother the average listener [2]. However, the causal relationship between these conditions and stress is not fully understood; stress may be both a direct consequence of tinnitus and decreased sound tolerance and an external factor in their development and severity [1,3–7]. The time-varying and situation-dependent nature of tinnitus and decreased sound tolerance may restrict the conclusions that can be drawn from laboratory studies, which are often limited in duration and external validity [8].

Recent advancements in biosignal-based stress recognition with wearables have offered the potential for automatically monitoring stress over longer periods of time in real-life settings [9]. Hearables, wearable devices worn at the level of the ear, could be especially well-suited for research on tinnitus and decreased sound tolerance as hearables are already commonly used to manage these conditions [2,5,10,11,11,12]. Given real-time information on the wearer's psychophysiological state and external environment, hearables may also be able to deliver improved therapies over time [13,14].

Heartbeat sounds, amplified inside the earcanal when wearing an occluding device, are among the biosignals that can be captured by a hearable and detected automatically [15,16]. Heartbeat signals are widely used in stress research [9], and previously, automatic stress and emotion recognition has been achieved with only heartbeat signals [17–25]. However, features derived from different types of heartbeat signals are not identical [26] and can be sensitive to artifacts common in wearable sensor data [27]. While stress monitoring with electrocardiography (ECG) signals has been well-established [9], it is currently

* Corresponding author at: École de technologie supérieure, 1100 Notre-Dame St W, Montreal, H3C 1K3, Quebec, Canada.
E-mail address: rachel.bouserhal@etsmtl.ca (R.E. Bouserhal).

https://doi.org/10.1016/j.compbiomed.2024.109555
Received 23 March 2024; Received in revised form 4 December 2024; Accepted 6 December 2024
Available online 31 December 2024

unclear whether similar results could be achieved with automatically processed in-ear heartbeat sounds. Therefore, the aim of this work is to assess the feasibility of stress monitoring using heartbeat sounds recorded with an in-ear device, by evaluating a stress recognition system to predict acute stress induced by experimental tasks in a typical population.

Contributions. To this end, a dataset was created with clinical-grade ECG and in-ear audio signals collected from 30 healthy young adults when they were at rest and performing two stress tasks: mental arithmetic [28] and the Cold Pressor Test [29]. This dataset was used to train systems for stress recognition. These systems were first benchmarked using heart rate variability (HRV) features extracted from the ECG signals. Then, the same features were extracted from audio signals captured within the occluded earcanal, and the performance of the resulting classifiers were compared to those trained on the ECG. Finally, as the error between features extracted from ECG and in-ear audio was higher during the stress conditions, a data augmentation method was designed to control for this error, and its utility was empirically verified.

2. Background

Psychological stress can affect the way the autonomic nervous system regulates bodily functions such as heart activity, respiration, and perspiration [9]. It has been suggested that tinnitus and decreased sound tolerance involve activation of the autonomic nervous system [5], and measurements of autonomic nervous system-dependent biosignals have been used to study these conditions [30–37]. In the general context of ambulatory stress monitoring, automatic stress recognition systems have been developed using biosensors, with some multimodal systems relying on multiple biosignal inputs [38–40], while others have been developed using only one type of sensor [17–21,24,25].

Heartbeat signals have shown promise for measuring stress, as both the parasympathetic "rest and digest" and sympathetic "fight or flight" responses of the autonomic nervous system influence heart activity [41]. Beat-to-beat variability in the heart rate can interact with different bodily functions such as respiration and blood pressure [42]. These systems can influence the heart rate simultaneously in different ways, and changes in the heart rate at different time scales can reflect separate neurophysiological mechanisms [43]. Given the exact number of milliseconds between consecutive heartbeats, referred to as interbeat intervals, multiple HRV measures can be extracted. These measures can be broadly categorized into time-domain, frequency-domain, and nonlinear features depending on the signal processing methodology used to extract them from the interbeat interval signal [42]. ECG is considered to be the gold standard heartbeat signal for HRV analysis, as ECG recordings are less likely to be distorted by artifacts and because the interbeat intervals can be precisely computed from the difference between the sharp, spike-like R-peaks in the ECG signal [44]. Accordingly, much of the research on automatic stress recognition to date has focused on HRV features extracted from ECG [9,17–21,24,25]. Nevertheless, in recent years, numerous novel data acquisition methods have been proposed for measuring HRV, such as contactless radar [45], wristband photoplethysmogram (PPG) [46], and audio signals recorded from the shoulder [23].

3. Material and methods

3.1. Dataset

3.1.1. Participants

In order to first establish the feasibility of stress monitoring with an in-ear device in a typical population, 30 healthy and normal hearing young adults were included in this study (15 men, age 27.3 ± 4.7 years and 15 women, age 27.1 ± 3.1 years). Recruitment was conducted through advertisements within the Université de Montréal, social networks, and word-of-mouth. To be included in the study, participants had to be aged between 18 and 45 years old and consider themselves to be physically and psychologically healthy. Exclusion criteria were having hearing disorders including tinnitus or hyperacusis, cardiovascular and cerebrovascular disorders, respiratory diseases, Raynaud's syndrome, diabetes, hyperglycemia, or taking any medication that could interfere with stress reactivity. Each participant signed a written informed consent before starting the experiment, and the project was approved by the Institutional Review Boards of Université de Montréal and École de technologie supérieure (QC, Canada) on March 9, 2020 (reference: H20200207).

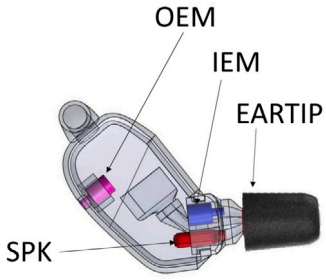


Fig. 1. Diagram of an Auditory Research Platform earpiece, containing an outer-ear microphone (OEM), an in-ear microphone (IEM), an internal miniature loudspeaker (SPK), and a high-attenuation foam eartip.

3.1.2. Materials

Hearing thresholds. Audiometry was performed in a soundproof audiometric booth using a calibrated AC40 clinical audiometer (Interacoustics, Middelfart, Denmark) with Telephonics TDH-39 headphones. Thresholds were measured with pure tones in octave bands between 250 Hz and 8,000 Hz using the Hughson-Westlake auditory threshold tracking procedure. Participants' eligibility was determined based on having a pure tone average (PTA) no greater than 15 dB HL (i.e., within the normal hearing limits) at frequencies of 0.5, 1, 2 and 4 kHz.

Auditory Research Platform In-ear heartbeat sounds were recorded with the Auditory Research Platform (ARP), an in-ear wearable technology developed within the ÉTS-EERS Industrial Research Chair in In-Ear Technologies (CRITIAS). Each earpiece, illustrated in Fig. 1, contains an outer-ear microphone (OEM), an in-ear microphone (IEM), an internal miniature loudspeaker (SPK), and a Comply Professional Noise Isolating eartip (Hearing Components, North Oakdale, MN, USA). Four channels of audio were recorded from the two OEMs and IEMs at a sampling rate of 44,100 Hz with MATLAB 2020a (Mathworks, Natick, MA, USA). The right earpiece was in "transparency mode", meaning that external sound picked-up by the OEM was played back on the SPK inside the wearer's ear at unity gain. Only the IEM signal acquired from the left earpiece was used for the experiments described in this paper.

Electrocardiogram (ECG) A clinical Burdick Altair-disc Holter (Spacelabs, Deerfield, USA) with five Ag/AgCl self-adhesive electrodes (3M Healthcare, Canada) positioned in 5-lead configuration was used for continuous heart rate recording. Raw ECG signal was sampled at a rate of 500 Hz.

3.1.3. Procedure

Once inclusion criteria were verified, participants were asked to enter a double-wall audiometric booth where the experiment took place, and general instructions about the experiment were given. Participants were then equipped with the Holter and ARP, which were respectively recording cardiac and audio signals simultaneously and continuously until the end of the experiment. The proper insertion of the ARP earpieces was checked with a fit-test [47]. Then, participants began the

stress measurement protocol, which consisted of three stress tasks in a pre-defined order (1. mental task, 2. noise exposure, and 3. cold pressor test) and 5-minute rest periods (sitting in silence) before each stress task. This order was chosen to allow participants to practice the mental task shortly before the start and to avoid potential lingering pain effects from the cold pressor test influencing subsequent tasks or rest periods. Specific instructions about the tasks were given by the experimenter to participants directly before each task started.

Mental arithmetic task Mental arithmetic tasks have been widely used to experimentally induce acute stress [21,28,39,48–51]. A stress task based on the Trier Social Stress Test (TSST) [28] was adapted to be performed silently, since the participant's voice, amplified by the occluded earcanal [52], would have otherwise contaminated the in-ear heartbeat sounds recorded by the ARP. Moreover, speech is also known to modify the breathing pattern, and consequently, the heart rate [53]. The task was developed on MATLAB and displayed on a PC screen using PsychToolBox library functions. Participants were instructed to count down from 1022 with a decrease of 13 (hence reported values of 1009, 996, etc.) until the five-minute task was over. The socio-evaluative component usually present in the TSST was replaced in this task by a visual timer limiting response time to 7.5 s. Additionally, participants received positive or negative feedback depending on the answer's accuracy. When a wrong or no answer was given, the participant had to restart the equations from the beginning (i.e., subtracting 13 from 1022). Regardless of how frequently the task was restarted, the total duration was limited to five minutes.

Noise task A broadband noise stimulus based on the study by Waye et al. [54] was generated with the Audacity audio editing software (www.audacityteam.org). This noise was chosen as it had previously been used on a population with noise sensitivity and tinnitus [54,55]. The stimulus magnitude was of rising intensity (ramped from 40 to 90 dBA) for 2 min and kept constant at 90 dBA for the remaining 3 min of the 5 min stimulus. The rising ramp was intended to be unpredictable for the participant, which is believed to induce stress [56,57]. The noise was sent through the AC40 audiometer, and played on two free field speakers inside the audiometric booth, oriented at 45° on either side of the participant's head. This task was included in the protocol to evaluate its ability to induce stress in a separate analysis, since noise is thought to be implicated in triggering or modulating tinnitus [58,59]. However, as this task has not been well-validated as a stressor, it was not intended to serve as ground truth stress data. Consequently, recordings from this noise task were not used as labeled data for training the stress recognition system, as further described in 3.2.1.

Cold Pressor Test During the cold pressor test, a standard pain task commonly used in pain and stress research [4,39,51,60], participants immersed one hand in cold water for 3 min. The water was contained in a cooler and the temperature was kept at 6.5 °C thanks to a thermal controller of 0.1 °C-sensitivity. A recirculating pump was installed inside the cooler to ensure water-circulation and thus preventing local warming of water temperature around the immersed hand.

3.2. Preprocessing

3.2.1. Task segmentation

The time points corresponding to each step of the experimental protocol were labeled by the experimenter using the Audacity software. In order to have a uniform segment duration across tasks and participants, for each task and rest period, a single three-minute segment was chosen. While the recommended segment duration for short-term HRV analysis is five minutes [9,42,44,61], some HRV features have been estimated with shorter durations [42,44]. Furthermore, shorter segment durations are commonly used for stress prediction, in particular when real-time analysis is envisioned [17,20,38–40,62–65].

The three-minute segment used for the rest period was chosen to maximize the time between steps in the protocol (i.e. when the last task ended and when preparation for the next task began). Though the participants were at rest for longer than three minutes, discarding the data from the start and end of the rest period accounted for recovery from the previous task and anticipation of the next task, as "residual" stress induced by experimental tasks may linger during the rest period [39,51]. The segment used for the tasks began 30 s before the start of each task and ended 150 s after the start. The timing was chosen to include data recorded while participants were anticipating the task [which can be a stressor on its own, see e.g. 66], as well as during the task itself.

For each participant, four three-minute segments were used to train and evaluate a binary stress classification model: two segments recorded during the mental arithmetic and cold pressor tasks were labeled as "stress", and two segments recorded during the rest periods prior to these two tasks were labeled as "rest". As the participants did not provide any subjective ratings of stress, the ground truth stress labels were solely based on the current experimental task. The segment recorded during the noise task was not used to train the stress classification model, as it has not been well-validated as a stress-induction task and may have been perceived differently by participants while they were wearing the earpiece. However, both the segments recorded prior to and during the noise task were used for synthesizing features with error, which is further explained in Section 3.3.2. Additionally, the rest segment recorded prior to the noise task was used for baseline normalization, further explained in Section 3.3.3.

3.2.2. Synchronization

The ECG and IEM data were synchronized as follows: at each step of the experimental protocol, an event marker button on the Holter monitor was pressed, triggering a sound that was recorded by the outer-ear microphones of the Auditory Research Platform and saving the corresponding timestamp on the Holter. The event marker sounds were labeled in Audacity by the experimenter. The Holter and Auditory Research Platform data were then aligned such that the mean difference between all Audacity labels and Holter event times was minimized. After synchronization, the maximum absolute difference between individual Audacity labels and Holter event times across all participants was 3.59 s. This range of synchronization error was considered acceptable for the current analysis which used three-minute long segments.

3.2.3. Heartbeat annotations

The ECG R-Peaks were visually inspected and corrected for artifact and ectopy using Burdick Vision Premier Holter analysis software (Cardiac Science-Quinton-Burdick, Bothell, WA, USA) and MATLAB. Only sinus beats were included in the computation of interbeat intervals. A small number of interbeat intervals were missing due to poor signal quality. However, as the total duration of missing intervals was never longer than 25% of the total segment length, none of the segments were excluded for having too much missing data [67].

The peaks corresponding to the first heartbeat sound in the IEM were automatically annotated following the methods described by Martin and Voix [15]. There was no manual correction of the IEM annotations, as some errors would be expected in a real-world monitoring context and it was of interest to assess how the stress classification would be affected by these errors.

3.3. Prediction

3.3.1. Feature extraction

HRV features were extracted using the Python toolbox Neurokit2 [68]. All time-domain, frequency-domain, and non-linear features available in Neurokit2 were used, excluding features that had invalid values; for example the standard deviation of the average of normal-to-normal intervals over two minutes (SDANN2) requires at least three two-minute windows, which was not possible with the three-minute segment length used in this analysis. A description of all 75 features used can be found in the supplemental material.

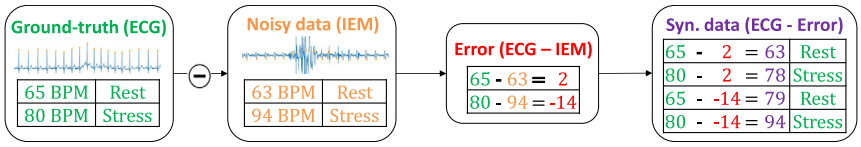


Fig. 2. Diagram of synthetic data generated with error for one feature and two samples. Note that for illustrative purposes, heart rate expressed in beats per minute (BPM) is given as an example feature, though this was not actually a feature used in this study (the mean and median heart period were in milliseconds).

There was no empirical selection of features using the current dataset, however, the stress classification performance was also evaluated using two smaller feature sets selected based on previous work for improved interpretability. The first feature set consisted of only the median heart period (MedianNN), providing information on the central tendency of the interbeat interval distribution. As lower values of MedianNN indicate a higher heart rate, decreased MedianNN is considered an indicator of "fight or flight" sympathetic activity [69]. MedianNN was chosen as a feature because it was expected to be relatively robust to errors in the automatic heartbeat annotation and it has previously been used for stress recognition [24,25,39,51]. The second feature set consisted of the MedianNN and the root-mean-square of successive differences (RMSSD). RMSSD provides information on the dispersion of the interbeat interval distribution. As it is based on the differences between successive beats, RMSSD indicates short-term changes in the interbeat intervals, which primarily reflect "rest and digest" parasympathetic activity [42]. RMSSD is one of the most commonly used HRV features and is well-suited for shorter segment durations [43].

3.3.2. Error-balanced data synthesis

A variety of sounds beyond the heartbeat can be amplified in an occluded earcanal [70]. Some of these sounds, such as noise artifacts caused by the wearer moving or swallowing, may induce errors in the automatic extraction of heartbeats from in-ear audio [15]. Heartbeat signal artifacts can distort heartbeat variability features, affecting the predictions of classifiers trained on these features [27].

Previous research using another type of wearable sensor to monitor stress (a wristband photoplethysmogram) has found that performing experimental tasks, regardless of whether they are intended to induce stress, can affect signal quality and reduce the number of accurately detected heartbeats compared to rest recordings [46]. In this work, the only "non-stress" class corresponded to the rest recordings, during which the participants were asked to sit in silence and do nothing rather than to perform an experimental task. Therefore, there is a risk that a classification system could learn task-related artifacts as a way to discriminate between rest and stress conditions, rather than physiologically meaningful differences between these conditions (for an empirical justification of this claim see Section 4.2).

To address this risk, a synthetic dataset containing equal errors across classes was generated, such that the learned classifier could be robust to the errors caused by artifacts in the IEM data without using these errors as a way to discriminate between classes. The method used for the synthetic data generation is presented in Algorithm 1, and a diagram is shown in Fig. 2.

Data synthesis was applied to the current dataset as follows: First, the error was computed by subtracting each HRV feature extracted from IEM from each feature extracted from simultaneously recorded ECG. This process resulted in one vector of feature errors per sample. Synthetic data was separately generated for each participant, and therefore, the errors from the test set were not used in generating the training data (the classifiers were trained and tested on different groups of participants, as further explained in Section 3.3.5). For the purposes of obtaining the error, the stress condition was considered to be irrelevant, and therefore, all six segments recorded during the three rest periods and the three experimental tasks (mental, noise and cold) were used to compute the feature errors. These six feature error vectors were then individually subtracted from the four ECG feature vectors to be used for stress prediction (corresponding to the two rest periods and the mental and cold tasks), resulting in 24 synthetic vectors per participant. Given that there were 30 participants and two stress tasks, there were 60 samples per class for the IEM and ECG. The synthetic data therefore consists of 360 samples per class (60 samples multiplied by the six feature error vectors per participant). The total number of samples per class for each signal is summarized in Table 1.

Table 1
Total number of samples per class for each type of signal.

	IEM	ECG	SYN
Stress	60	60	360
Rest	60	60	360

3.3.3. Feature normalization

Normalization of heart rate variability features can be achieved at different steps in the prediction pipeline: the interbeat intervals can be normalized prior to HRV feature extraction [39,63], or the HRV features themselves can be normalized [17,46,65,71,72,72]. As certain features are invariant to scaling of the interbeat intervals, such as the ratio of low frequency to high frequency power (LFHF), normalization was always implemented at the HRV feature level such that all features would be transformed by each normalization step. Furthermore, normalization can be applied within participants [17,46,65,71,72], as well as using the statistics from participants in the training set [72]. In this work, two types of normalization were applied to the HRV features:

Baselining. Normalizing physiological signals according to each individual is a common feature transformation prior to stress prediction; however, it is unclear how baseline normalization would be implemented in real-world scenarios in which an individual's physiology would likely vary over longer periods of time [39]. Therefore, one aim of this work was to evaluate the feasibility of stress prediction without normalizing using an individual's baseline. When baselining was applied, each feature value for a participant was divided by the mean of that feature's values from all baseline periods for that participant [71]. Since rest periods were used as the non-stress class in stress classification and no separate baseline recordings were available, baseline feature values were extracted from the rest periods before the two tasks other than the current task. For example, features from both the rest period prior to the cold task and the cold task itself were normalized with the features from the rest periods prior to the mental and noise tasks. The two other rest periods were used instead of the three rest periods or the entire dataset in order to simulate baselining with features from separate baseline recordings.

Z-scoring. As a separate normalization step, each feature was z-scored with StandardScaler in the Scikit-learn pipeline [73]. Z-scoring was applied to both baseline-normalized features and features that were not baseline-normalized. The mean and standard deviation were computed based on the training data, that is, the samples corresponding to two rest and two stress tasks for all participants selected as the training set for the current iteration of the cross-validation (further explained in Section 3.3.5). All samples in both the training and test sets were z-scored using the mean and standard deviation derived from the training set. The z-scoring operation was the final feature transformation before classification, and it was applied in all experiments.

Algorithm 1 NumPy-style pseudocode for error-balanced data synthesis

```
def get_error(X_ecg, X_iem):
    error = X_ecg - X_iem
    return error

def synthesize_with_error(X_ecg, error, y):
    # N: number of samples used to compute error
    # D: dimensionality of the feature vector
    N, D = error.shape
    # subtract all error vectors
    # from every ECG sample
    X_syn = (X_ecg[:,None,:] - error[None,:,:])
    X_syn = X_syn.reshape(-1, D)
    # generate synthetic labels
    # by repeating labels corresponding to ECG
    y_syn = y.repeat(N)
    return X_syn, y_syn
```

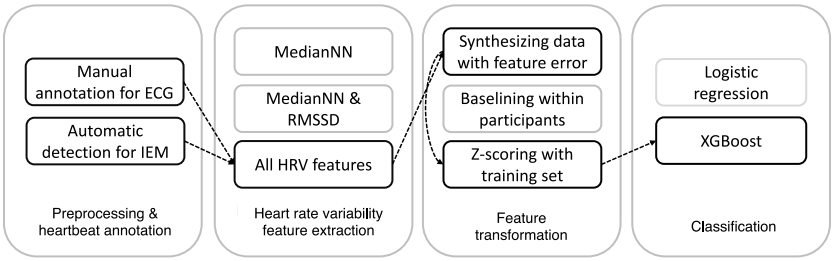


Fig. 3. Diagram depicting possible components of the prediction pipeline. Arrows indicate the components included in the pipeline developed in the final set of experiments.

3.3.4. Classification algorithms

Prior work has proposed automatic stress recognition systems based on a wide variety of classification methods ranging in complexity and interpretability, including linear models, ensemble learning, and deep learning [9]. In this work, two classification algorithms were tested: logistic regression, implemented in Scikit-learn [73], and extreme gradient boosting (XGBoost), implemented in the XGBoost Python package [74]. Logistic regression is a simple linear model that has been previously used to classify heartbeat data [20,25,38,71,75–77], with the advantage of being intrinsically interpretable [78]. On the other hand, XGBoost is a more complex ensemble model, with previously demonstrated performance classifying various types of data including heartbeat data [62,75,77,79,80]. The default hyperparameters for both of these algorithms were used except that the maximum number of iterations was set to 1000 rather than 100 for logistic regression.

3.3.5. Evaluation

Different combinations of heartbeat signals, HRV features, feature transformations, and classifiers (depicted in Fig. 3) were tested in a series of experiments. In all experiments, each stress recognition system was evaluated with a repeated five-fold participant-wise cross-validation scheme, in order to test how the system would generalize to unseen individuals. Participants were randomly divided into five groups, and the samples from participants in four of the five groups were used to train the classifier, rotating which group was used to test the classifier. This procedure was repeated with 10 random seeds, resulting in 50 evaluations of each system tested (with the same 10 random seeds used for each system). Performance was assessed with the mean and standard deviation of the accuracy over these 50 evaluations.

The performance of the stress recognition systems was evaluated in several analyses guided by the following questions:

1. How does normalizing to each individual's baseline affect stress recognition performance?
2. How does the performance of stress recognition systems trained on in-ear audio compare to those trained on clinical-grade ECG?
3. How does performance change when using a synthetic dataset in which the feature error is balanced across rest and stress conditions?

4. Results

4.1. Baseline normalization

To assess the upper bound of stress classification performance with a heartbeat signal with baselining and without, classifiers trained and tested on the ECG were compared when the features were baselined vs. when their original values were used. Additionally, the two classifiers, as well as the three feature sets, were compared. The accuracy of classifiers trained on the ECG and tested on the ECG with baselining is presented in Table 2. With baselining, the highest mean accuracy achieved was 78.9% using all HRV features; however, reducing the feature set had a relatively small impact on performance: the highest mean accuracy using MedianNN & RMSSD was 76.7% and the highest mean accuracy using only MedianNN was 76.0%. In contrast, without baselining, using the two smaller feature sets led to a larger drop in performance: compared to 76.5% with all HRV features, the highest accuracy using MedianNN & RMSSD was 59.9%, and the highest accuracy using only MedianNN was 63.7%. The accuracy of all classifiers trained on the ECG and tested on the ECG without baselining is listed in Table 3.

4.2. Comparison of ECG and in-ear audio

After the feasibility of stress classification without baselining was successfully established with the ECG, it was of interest to evaluate whether a satisfactory performance could be achieved even when using a less reliable heartbeat signal, extracted from the IEM audio. To this end, the same combinations of feature sets and classifiers were trained and tested with HRV features extracted from the IEM data. The performance of these models is presented in Table 4. Surprisingly, the models

Table 2
Accuracy of stress classifiers trained on the ECG and tested on the ECG with baselining.

Features	Classifier	Acc. (%)
MedianNN	Log. Reg.	76.0 ± 7.3
MedianNN	XGBoost	67.5 ± 8.4
MedianNN & RMSSD	Log. Reg.	76.7 ± 8.5
MedianNN & RMSSD	XGBoost	72.0 ± 9.4
All HRV features	Log. Reg.	**78.9 ± 8.1**
All HRV features	XGBoost	76.3 ± 7.8

Table 3
Accuracy of stress classifiers trained on the ECG and tested on the ECG without baselining.

Features	Classifier	Acc. (%)
MedianNN	Log. Reg.	57.2 ± 5.8
MedianNN	XGBoost	63.7 ± 7.1
MedianNN & RMSSD	Log. Reg.	59.5 ± 6.1
MedianNN & RMSSD	XGBoost	59.9 ± 7.9
All HRV features	Log. Reg.	**76.5 ± 8.4**
All HRV features	XGBoost	73.0 ± 8.5

Table 4
Accuracy of stress classifiers trained on the IEM and tested on the IEM without baselining.

Features	Classifier	Acc. (%)
MedianNN	Log. Reg.	62.2 ± 6.1
MedianNN	XGBoost	58.7 ± 10.3
MedianNN & RMSSD	Log. Reg.	75.9 ± 6.5
MedianNN & RMSSD	XGBoost	79.1 ± 7.1
All HRV features	Log. Reg.	**81.1 ± 8.0**
All HRV features	XGBoost	76.1 ± 6.7

trained and tested on the IEM had higher accuracy than the models trained and tested on the ECG for the feature sets containing MedianNN & RMSSD and all HRV features. When training and testing on the IEM, the model with the highest mean accuracy without baselining was Logistic Regression using all HRV features: 81.1% (Table 4), compared to 76.5% (Table 3) when training and testing with ECG.

A substantial difference in accuracy between the classifiers trained and tested on the IEM vs. ECG was observed for the MedianNN & RMSSD feature set. For this feature set, the best-performing model trained and tested on the IEM had a mean accuracy of 79.1% (Table 4), while the best-performing model trained and tested on the ECG had a mean accuracy of 59.9% (Table 3). With MedianNN as the only feature, the performance was more similar between the classifiers trained and tested on the IEM (62.2%; Table 4) vs. ECG (63.7%; Table 3).

To assess whether the models trained on HRV features extracted from the IEM could generalize to conventional HRV features, all models trained on the IEM were tested on the ECG. The classification performance is presented in Table 5. With the feature set consisting of only MedianNN, there was a relatively small reduction in mean accuracy when comparing the best-performing model trained on the IEM and tested on the ECG (62.7%, Table 5) to the best-performing model both trained and tested on the ECG (63.7%, Table 3). However, with the feature sets containing MedianNN & RMSSD and all HRV features, mean accuracy training on the IEM and testing on the ECG ranged from 50.0% to 50.8%.

The coefficients of the logistic regression models separately trained on the ECG and IEM, each using the MedianNN & RMSSD feature set, are shown in Fig. 4. For both the models trained on ECG and IEM, the coefficients for MedianNN were negative, indicating an increased probability of predicting stress when MedianNN was lower, and the coefficients for RMSSD are positive, indicating an increased probability of predicting stress when RMSSD was higher. As the feature values were z-scored with the mean and standard deviation of all the samples in the training set, the magnitude of the coefficient for each feature was considered to be a measure of importance [76]. In terms of which of

Table 5
Accuracy of stress classifiers trained on the IEM and tested on the ECG without baselining.

Features	Classifier	Acc. (%)
MedianNN	Log. Reg.	57.8 ± 7.5
MedianNN	XGBoost	**62.7 ± 8.5**
MedianNN & RMSSD	Log. Reg.	50.8 ± 1.7
MedianNN & RMSSD	XGBoost	50.1 ± 0.6
All HRV features	Log. Reg.	50.0 ± 0.0
All HRV features	XGBoost	50.2 ± 0.8

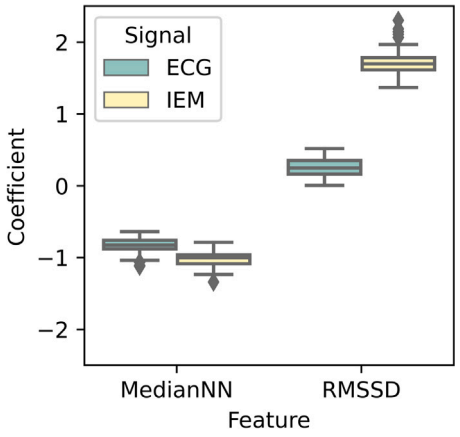


Fig. 4. Box plot of feature coefficients in logistic regression models trained on the ECG and IEM. The coefficients were derived from models using the MedianNN & RMSSD feature set without baselining, trained over all cross-validation iterations. MedianNN had a higher absolute coefficient value than RMSSD in the model trained on the ECG, while RMSSD had a higher absolute coefficient value than MedianNN in the model trained on the IEM. Note that a negative coefficient indicates increasing probability of predicting stress when the standardized feature values are negative, while a positive coefficient indicates increasing probability of predicting stress when the standardized feature values are positive.

the two features was more important, the models trained on the two signals had opposite trends. In the case of the model trained on the ECG, MedianNN (mean coefficient of −0.8) had higher importance than RMSSD (mean coefficient of 0.3). In contrast, when training on the IEM, MedianNN (mean coefficient of −1.0) had lower importance than RMSSD (mean coefficient of 1.7). As demonstrated in Fig. 4, this pattern was consistent across cross-validation iterations.

Aiming to understand the source of the difference in feature importance, the error between HRV features extracted from ECG and from IEM was assessed. Unlike the manually-corrected ECG annotations, the heartbeats in the IEM data were annotated automatically (without manual correction of missed or extra beats) such that the IEM annotations would be representative of a real-world monitoring context. After manual review of segments of the in-ear audio where the HRV feature error was high, it became apparent that the IEM signal contained many artifacts such as movement and speech, despite the fact that participants were instructed to remain still and silent during the experimental tasks. An example of an IEM signal with artifacts and the corresponding heartbeat annotation is shown in Fig. 5. As some of these artifacts may have been related to the task (e.g. keyboard sounds during the computer-based mental task), the error was assessed separately for the rest and stress conditions.

The agreement between ECG and IEM for MedianNN and RMSSD was visually assessed with Bland-Altman plots, shown in Fig. 6. Visual inspection of the plots suggests that MedianNN extracted from IEM data in some cases overestimated and in other cases underestimated MedianNN extracted from ECG data (though the mean difference was negative). In contrast, the plots demonstrate that RMSSD extracted from IEM data tended to consistently overestimate RMSSD extracted from

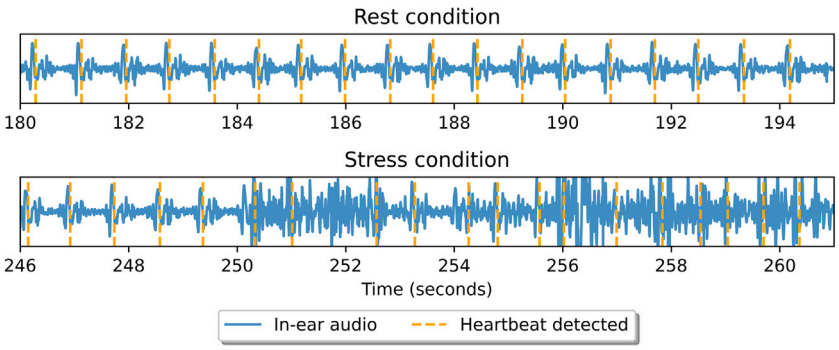
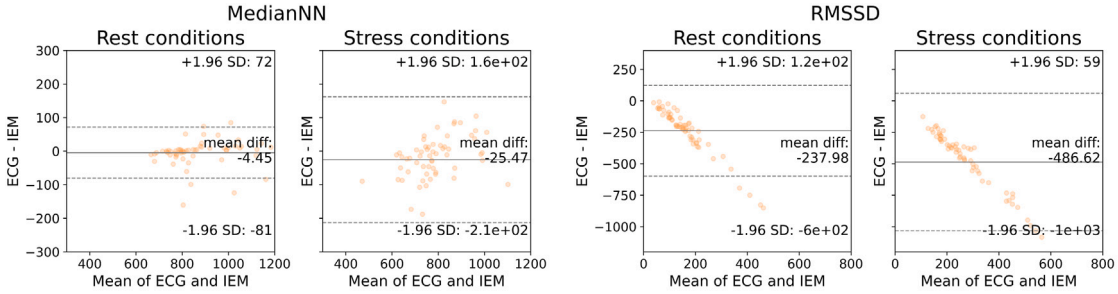


Fig. 5. Example IEM recording and automatically detected heartbeats during a rest and stress condition. It can be seen that when there were artifacts in the IEM signal, the heartbeats were misdetected.



(a) Agreement for MedianNN (b) Agreement for RMSSD

Fig. 6. Bland-Altman plots assessing agreement between the electrocardiogram (ECG) and in-ear microphone (IEM) on the HRV features (a) MedianNN and (b) RMSSD, for data recorded during rest and stress conditions. Each point represents the difference between a HRV feature extracted from ECG and the same feature extracted from simultaneously recorded IEM data. The solid line in the middle represents the mean difference, while the upper and lower dashed lines represent the upper and lower limits of agreement at 95% confidence.

ECG data, particularly in the stress conditions: the mean difference is −486.53 in stress conditions but is −237.99 in rest conditions. In both rest and stress conditions, it can be seen that as RMSSD extracted from IEM increasingly overestimated RMSSD extracted from ECG (shown on the y-axis), the mean of the RMSSD values extracted from ECG and IEM (shown on the x-axis) tended to increase, suggesting that the error greatly influenced the value of RMSSD for the IEM. The supplemental material includes further analyses comparing ECG and IEM for additional HRV features that were important for the best-performing model trained and tested on IEM.

4.3. Error-balanced data synthesis

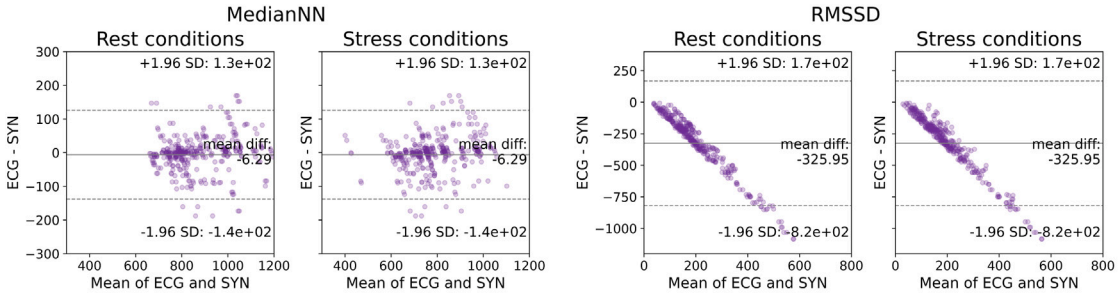
After it was determined that the HRV feature error differed in the rest and stress conditions, synthetic data (SYN) was generated to balance the errors across classes, following the methods described in Section 3.3.2. The agreement between ECG and SYN for MedianNN and RMSSD was assessed with Bland-Altman plots, shown in Fig. 7. When comparing to the Bland-Altman plots assessing agreement between ECG and IEM (Fig. 6), it can be seen that the distribution of the synthetic errors between ECG and SYN in each of the rest and stress conditions is similar to the distribution of the original errors between ECG and IEM: SYN MedianNN sometimes overestimated and sometimes underestimated ECG MedianNN, while SYN RMSSD largely overestimated ECG RMSSD. However, unlike the original errors between ECG and IEM, the mean difference between ECG and SYN is identical across classes, for MediannNN (mean difference of −6.29) and RMSSD (mean difference of −325.92).

All models trained on the IEM were tested on the SYN, in order to evaluate whether the models trained on HRV features extracted from the IEM would perform as well when tested on data where the overall distribution of the errors was similar to IEM but where these errors were balanced across classes (and therefore could not be used as a way to discriminate between rest and stress). The classification performance of the models trained on the IEM and tested on the SYN is presented in Table 6. With the feature set consisting of only MedianNN, the mean accuracy was similar when comparing the best-performing model trained on the IEM and tested on the SYN (61.1%, in Table 6) to the best-performing model both trained and tested on the IEM (62.2%, in Table 4). With the feature sets containing MedianNN & RMSSD and all HRV features, there was a larger drop in performance: mean accuracy training on the IEM and testing on the SYN ranged from 54.0% to 64.9% (see Table 6), compared to a mean accuracy ranging from 75.9% to 81.1% when both training and testing on the IEM (see Table 4).

As the performance of models trained on the IEM decreased when the error was balanced across classes compared to testing on the IEM, it was of interest to assess whether augmenting the training data with the SYN would allow for sufficient performance when testing on IEM without the possibility of learning class-dependent differences in errors. To determine the added benefit of augmenting the training data with errors, models were first trained on just the clean (manually-corrected) ECG and tested on the noisy (automatically-annotated) IEM. The classification performance of these models is presented in Table 7.

Table 6
Accuracy of stress classifiers trained on the IEM and tested on the SYN without baselining.

Features	Classifier	Acc. (%)
MedianNN	Log. Reg.	61.1 ± 5.4
MedianNN	XGBoost	59.7 ± 4.7
MedianNN & RMSSD	Log. Reg.	59.1 ± 4.5
MedianNN & RMSSD	XGBoost	54.0 ± 2.3
All HRV features	Log. Reg.	**64.9 ± 5.3**
All HRV features	XGBoost	60.9 ± 5.0



(a) Agreement for MedianNN (b) Agreement for RMSSD

Fig. 7. Bland-Altman plots assessing agreement between the electrocardiogram (ECG) and synthetic data (SYN) on the HRV features (a) MedianNN and (b) RMSSD, for data recorded during rest and stress conditions. Each point represents the difference between a HRV feature extracted from ECG and the same feature synthesized with one error value and the original ECG feature.

Table 7
Accuracy of stress classifiers trained on the ECG and tested on the IEM without baselining.

Features	Classifier	Acc. (%)
MedianNN	Log. Reg.	61.3 ± 6.3
MedianNN	XGBoost	60.8 ± 9.5
MedianNN & RMSSD	Log. Reg.	54.2 ± 4.9
MedianNN & RMSSD	XGBoost	56.0 ± 9.2
All HRV features	Log. Reg.	**65.4 ± 9.3**
All HRV features	XGBoost	60.7 ± 7.3

Table 8
Accuracy of stress classifiers trained on the ECG & SYN using all HRV features without baselining.

Test signal	Classifier	Acc. (%)
ECG	Log. Reg.	74.3 ± 8.1
ECG	XGBoost	**76.7 ± 7.9**
IEM	Log. Reg.	69.9 ± 9.0
IEM	XGBoost	**76.6 ± 7.0**
SYN	Log. Reg.	70.6 ± 5.8
SYN	XGBoost	**77.3 ± 4.7**

The best-performing model, logistic regression trained on all HRV features had a mean accuracy of 65.4%, still well above chance accuracy but a substantial drop from the same model trained and tested on the ECG (76.5%, Table 3). The same models were then trained on a dataset consisting of both ECG and SYN, and tested separately on ECG, IEM, and SYN to evaluate how augmenting the dataset would affect generalization to features extracted from clean data without error, real data with class-imbalanced errors, and synthetic data with class-balanced errors. The performance of the models trained using the best-performing feature set, all HRV features, is shown in Table 8. XGBoost had the highest mean accuracy for all test signals, ranging from 76.6% to 77.3%.

5. Discussion

5.1. Baseline normalization

The aim of this work was to determine the feasibility of stress monitoring with an in-ear wearable device. Since the features extracted from heartbeat sounds from an in-ear wearable device were expected to be noisier than the gold-standard ECG signals generally used for measuring HRV, the first set of analyses aimed to establish the upper bound of stress classification performance possible, using a conventional sensor. Performance ranged from 67.5% to 78.9%, within the range of performance previously achieved for stress classification with biosignals [9].

As normalization for each individual's baseline may be difficult to implement in a real-world monitoring context [39], the effect of baselining on classification performance using ECG was assessed. It was found that normalizing to each individual's baseline improved classification results, in line with previous research using HRV [72]. However, using all HRV features allowed for nearly as high performance as without baselining, suggesting that concerns about the ecological validity of the baselining procedure could be avoided by including these additional features.

Nonetheless, fewer features may be preferred for certain applications, as the number of features plays a major role in the interpretability of a model [81], which may be especially important in clinical contexts [82] such as treatment or diagnosis of tinnitus and decreased sound tolerance. Moreover, fewer features may be preferred to reduce computational cost [62,64], as hearables, like other wearables, generally have limited computational power. However, it may be possible that a smaller subset of the 75 HRV features used in this work could achieve similar performance without baselining, as many HRV features are correlated with each other and influenced by the same physiological phenomena [42,83]. Future efforts could aim to identify this smaller subset of features, ideally with a separate dataset for testing [39], as empirically selecting this subset from all the subsets possible given 75 features would require many comparisons, and therefore, there is an increased risk that improvements in mean cross-validation accuracy would not translate to better performance on completely unseen data [84].

5.2. Comparison of ECG and in-ear audio

A subsequent set of experiments compared stress classification systems trained on features extracted from the clean ECG signal to those trained on the features extracted from the IEM audio signal, to determine whether automatically-detected heartbeats in the IEM could estimate HRV well enough to predict stress. Unexpectedly, the best-performing classifier trained on the IEM performed better than the best-performing classifier trained on the manually-annotated, clinical-grade ECG. For models trained on just two features, MedianNN and RMSSD, training and testing with the IEM instead of the ECG improved the mean accuracy by 16 to 19% depending on the classifier used (compare Tables 3 and 4). However, analyses of the feature importance and error suggest that the improved performance of the models trained on the IEM was not due to learning physiologically meaningful features: RMSSD, the most important feature for one model trained on the IEM, was highly distorted by misdetected heartbeats and had a higher error during the stress conditions than during the rest conditions. This difference in error between conditions may have been due to artifacts more commonly found during the stress tasks such as the presence of movement, speech signals or keyboard typing noise, as these can increase the number of misdetected heartbeats.

Spurious correlations have been known to be an issue in deploying machine learning algorithms across different domains. Biases in the data used to train and evaluate a model can lead to large drops in performance once deployed [82]. For example, a pneumonia detection algorithm that appeared to perform well when tested on data from the same hospital did not generalize well to new hospitals, likely because the algorithm exploited confounding information specific to the hospital such as the type of scanner used in departments with different rates of pneumonia [85]. In the context of stress classification with heartbeat signals, it is also possible that spurious information in the training data can lead to large drops in classification performance when applied in unseen contexts, particularly when the input to a data-driven system is not limited to physiologically meaningful features based on domain knowledge: in one study, deep learning models directly trained on ECG signals appeared to perform better than models trained on HRV features when tested on held-out data from the same dataset used for training, but the HRV feature-based models outperformed the ECG signal-based models when tested on a different dataset [24]. The results of the current study suggest that even HRV features can be affected by spurious information if errors are not accounted for, and therefore, the cross-validation performance obtained by training and testing on HRV features extracted from the IEM likely does not reflect how well the model would perform when deployed. If the increased performance for the stress prediction is driven by the increased presence of artifacts during the stress tasks, the learned classifier would not generalize to real-world applications where artifacts such as movement, speech, and keyboard typing may occur when the wearer is not stressed. More broadly, these findings highlight the risk of artifacts unknowingly being exploited by predictive models and could have implications for any applications in which raw or automatically processed biosignals are used as input to a classifier without controlling for the presence of artifacts.

5.3. Error-balanced data synthesis

When the models trained on the IEM were evaluated with synthetic data in which the HRV feature errors were equal across classes, performance dropped substantially, further supporting the claim that the models trained on the IEM had relied on this difference in HRV feature error to discriminate between rest and stress. The influence of spurious patterns on predictions has been previously identified and reduced with synthetic data in other domains [86–88]. In the context of the natural language processing task of coreference resolution, imbalanced co-occurrence of gendered pronouns with certain occupations has been corrected for by generating new sentences with different pronouns (e.g. swapping “he” for “she” in “The physician hired the secretary because he was overwhelmed with clients”.). These modified sentences have revealed gender biases in data-driven coreference systems and shown to be an effective way to prevent biases when used as training data [88]. The current study extended the existing research on robustness to spurious patterns by proposing a data synthesis method to counter this issue for systems that learn from biosignals that may be corrupted by artifacts. While the analyses presented in this paper focused on the synthesis of HRV features, the proposed method could be applied to any dataset containing both physiologically meaningful “clean” features and “noisy” features from which errors can be computed (e.g. breathing rate derived from a conventional respiration sensor and breathing rate automatically extracted from in-ear audio signals [15]).

In addition to its utility to evaluate existing models for their reliance on spurious patterns, the synthesized data also proved useful for training: compared to training on only the clean ECG features, augmenting the dataset with the error substantially improved test performance on the IEM. Data augmentation methods that synthesize additional samples with noise are often employed to increase the number of samples, to improve performance of machine learning methods that tend to rely on a large amount of training data, as well as to make the system more robust to noise [89]. It may be possible to further improve the system’s robustness to noise at other steps in the prediction pipeline. When interpretability is crucial, it may be preferable to identify unreliable segments of the heartbeat signal before HRV feature extraction rather than have a classifier robust to errors in the HRV features. That said, if there are many of these unreliable segments, it may not be possible to have a long enough recording for HRV feature extraction without treating unreliable segments as missing data, and methods to treat missing data such as interpolation could still introduce errors in the HRV features [46,90].

One possible limitation in the theoretical assumptions made in this work should be noted: by using all of the errors to generate new data for both rest and stress conditions regardless of which condition the original errors came from, the proposed data synthesis method assumed that the difference between the ECG and the IEM should be completely independent of the stress condition. In reality, it is possible that the difference between the ECG and the IEM could also be influenced by factors related to the stress condition to some extent; for example, even without errors in the heartbeat annotations, there are small physiological differences in the interbeat intervals computed from different heartbeat signals, and these differences may constitute novel biomarkers [26]. Determining the extent to which physiological differences influence the error would be an interesting avenue for future research. However, the fact that the test performance was high for both IEM and SYN suggests that simply assuming that the error is completely independent of the stress condition is sufficient for the purposes of the data augmentation, that is, to achieve high performance on IEM data containing artifacts without unintentionally learning the artifacts as a way to discriminate between rest and stress.

Finally, future work could investigate the utility of data synthesis with separate datasets. In the current study, the same dataset was used to estimate the error and obtain the clean features that were augmented with the error. However, it is possible that the errors and the clean features could be derived from separate sources: the distribution of the feature errors could be modeled based on any dataset containing simultaneous clean and noisy features, and this distribution could be sampled from to augment a separate dataset containing only clean features recorded under the conditions of interest (e.g. stress tasks). The ability to augment separate datasets could save substantial resources associated with data collection, as it would open the possibility of developing algorithms for hearables with datasets that were not collected with a hearable in mind, only requiring in-ear audio signals to validate that the developed algorithms are robust to real errors.

6. Conclusions

This work demonstrated the feasibility of a novel modality of stress monitoring based on heartbeat sounds recorded from inside the human earcanal, which could lead to improved research methods and therapies by extending the capabilities of existing hearing technology for tinnitus and decreased sound tolerance. Moreover, the findings of this work highlighted the possibility that artifacts in biosignal data can be exploited by a classifier to achieve high performance that likely would not generalize to a real-world monitoring context. The data augmentation methodology presented in this work was shown to both effectively identify biases in models caused by artifacts and improve model robustness to artifacts, with implications for the broader field of automatic biosignal monitoring.

CRediT authorship contribution statement

Danielle Benesch: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Bérangère Villatte:** Writing – review & editing, Visualization, Formal analysis, Data curation. **Alain Vinet:** Writing – review & editing, Supervision,

Investigation, Funding acquisition, Conceptualization. **Sylvie Hébert:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Jérémie Voix:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization. **Rachel E. Bouserhal:** Writing – review & editing, Supervision, Resources, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank Ajin Tom, Philippe Chabot, and Gabrielle Crétot-Richert for their feedback on the initial analyses, Pascal Giard for his feedback on the manuscript, Omar Mahiddine for his expertise on the Holter and for the ECG signal inspection, and Valentin Pintat, Corentin Delain, and Frédérik Desaulniers for data acquisition setup and the initial measurements. This work was supported in part by the Natural Sciences and Engineering Council of Canada (NSERC), EERS Global Technologies Inc. and the 'ETS-EERS Industrial Research Chair in In-Ear Technologies (CRITIAS), Canada through the Alliance Grant (ALLRP 566678 - 21). This work was also made possible via funding from the FRQ AUDACE program, Canada (2020-AUDC-267674) and Agile Seed Funding from the Centre for Interdiscplinary Research in Music, Media and Technology (CIRMMT), Canada. The first author acknowledges financial support received from the German Academic Exchange Service (DAAD 57503736) and MITACS Accelerate Cluster, Canada (IT26677/SUBV-2021-168).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.109555.

Data availability

The implementation of the proposed model is available at https://github.com/critias-ca/in-ear-stress. Physiological data may be made available upon reasonable request. Requests can be sent via email to the corresponding author.

References

[1] B. Mazurek, B. Boecking, P. Brueggemann, Association between stress and tinnitus–New aspects, Otol. Neurotol. 40 (4) (2019) e467–e473, http://dx.doi.org/10.1097/MAO.0000000000002180.

[2] P. Jastreboff, M. Jastreboff, Treatments for decreased sound tolerance (hyperacusis and misophonia), Semin. Hear. 35 (02) (2014) 105–120, http://dx.doi.org/10.1055/s-0034-1372527.

[3] B. Mazurek, A. Szczepek, S. Hébert, Stress and tinnitus, HNO 63 (4) (2015) 258–265, http://dx.doi.org/10.1007/s00106-014-2973-7.

[4] D. Hasson, T. Theorell, J. Bergquist, B. Canlon, Acute stress induces hyperacusis in women with high levels of emotional exhaustion, PLoS One 8 (1) (2013) e52945, http://dx.doi.org/10.1371/journal.pone.0052945.

[5] P.J. Jastreboff, The neurophysiological model of tinnitus, in: Tinnitus: Theory Manag., B. C. Decker, 2004, p. 15.

[6] W. Schlee, R. Kraft, J. Schobel, B. Langguth, T. Probst, M.P. Lourenco, J. Simoes, P. Neff, R. Hannemann, M. Reichert, Momentary assessment of tinnitus–How smart mobile applications advance our understanding of tinnitus, in: Digital Phenotyping and Mobile Sensing, Springer, 2023, pp. 285–303, http://dx.doi.org/10.1007/978-3-030-98546-2_16.

[7] S. Hébert, B. Canlon, D. Hasson, Emotional exhaustion as a predictor of tinnitus, Psychother. Psychosom. 81 (5) (2012) 324–326, http://dx.doi.org/10.1159/000335043.

[8] F.H. Wilhelm, P. Grossman, Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment, Biol. Psychol. 84 (3) (2010) 552–569, http://dx.doi.org/10.1016/j.biopsycho.2010.01.017.

[9] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, M. Tsiknakis, Review on psychological stress detection using biosignals, IEEE Trans. Affect. Comput. (2019) http://dx.doi.org/10.1109/TAFFC.2019.2927337, 1–1.

[10] A.E. Cavanna, S. Seri, Misophonia: Current perspectives, Neuropsychiatr. Dis. Treat. (2015) 2117, http://dx.doi.org/10.2147/NDT.S81438.

[11] D. Neave-DiToro, A. Fuse, M. Bergen, Knowledge and awareness of ear protection devices for sound sensitivity by individuals with autism spectrum disorders, Lang., Speech, Hearing Serv. Schools 52 (1) (2021) 409–425, http://dx.doi.org/10.1044/2020_LSHSS-19-00119.

[12] C.A. Sammeth, D.A. Preves, W.T. Brandy, Hyperacusis: Case studies and evaluation of electronic loudness suppression devices as a treatment approach, Scand. Audiol. 29 (1) (2000) 28–36, http://dx.doi.org/10.1080/010503900424570.

[13] D. Benesch, K.N. Raj, R. Bouserhal, J. Voix, Interfacing the Tympan open-source hearing aid with an external computer for research on decreased sound tolerance, in: Proceedings of Meetings on Acoustics, vol. 45, Seattle, Washington, 2021, 050005, http://dx.doi.org/10.1121/2.0001616.

[14] G.D. Searchfield, P.J. Sanders, Z. Doborjeh, M. Doborjeh, R. Boldu, K. Sun, A. Barde, A state-of-art review of digital technologies for the next generation of tinnitus therapeutics, Front. Digit. Health 3 (2021) 724370, http://dx.doi.org/10.3389/fdgth.2021.724370.

[15] A. Martin, J. Voix, In-ear audio wearable: Measurement of heart and breathing rates for health and safety monitoring, IEEE Trans. Biomed. Eng. 65 (6) (2018) 1256–1263, http://dx.doi.org/10.1109/TBME.2017.2720463.

[16] X. Fan, D. Pearl, R. Howard, L. Shangguan, T. Thormundsson, APG: Audioplethysmography for cardiac monitoring in hearables, in: Proceedings of the 29th Annual International Conference on Mobile Computing and Networking, ACM, Madrid, Spain, 2023, pp. 1–15, http://dx.doi.org/10.1145/3570361.3613281.

[17] G. Giannakakis, K. Marias, M. Tsiknakis, A stress recognition system using HRV parameters and machine learning techniques, in: 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW, IEEE, Cambridge, United Kingdom, 2019, pp. 269–272, http://dx.doi.org/10.1109/ACIIW.2019.8925142.

[18] P. Melillo, M. Bracale, L. Pecchia, Nonlinear heart rate variability features for real-life stress detection. Case study: Students under stress due to university examination, BioMedical Engineering OnLine 10 (1) (2011) 96, http://dx.doi.org/10.1186/1475-925X-10-96.

[19] B. Szakonyi, I. Vassányi, E. Schumacher, I. Kósa, Efficient methods for acute stress detection using heart rate variability data from ambient assisted living sensors, BioMed. Eng. OnLine 20 (1) (2021) 73, http://dx.doi.org/10.1186/s12938-021-00911-6.

[20] B. Kaur, J.J. Durek, B.L. O'Kane, N. Tran, S. Moses, M. Luthra, V.N. Ikonomidou, Heart rate variability (HRV): An indicator of stress, in: H.H. Szu, L. Dai (Eds.), SPIE Sensing Technology + Applications, Baltimore, Maryland, USA, 2014, p. 91180V, http://dx.doi.org/10.1117/12.2051148.

[21] H.M. Cho, H. Park, S.Y. Dong, I. Youn, Ambulatory and laboratory stress detection based on raw electrocardiogram signals using a convolutional neural network, Sensors 19 (20) (2019) 4408, http://dx.doi.org/10.3390/s19204408.

[22] H. Ferdinando, L. Ye, T. Seppänen, E. Alasaarela, Emotion recognition by heart rate variability, Aust. J. Basic Appl. Sci. (2014) 7.

[23] C. Xiefeng, Y. Wang, S. Dai, P. Zhao, Q. Liu, Heart sound signals can be used for emotion recognition, Sci. Rep. 9 (1) (2019) 6486, http://dx.doi.org/10.1038/s41598-019-42826-2.

[24] P. Prajod, E. André, On the generalizability of ECG-based Stress Detection Models, in: 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), 2022, pp. 549–554, http://dx.doi.org/10.1109/ICMLA55696.2022.00090.

[25] K. Arquilla, A.K. Webb, A.P. Anderson, Utility of the full ECG waveform for stress classification, Sensors 22 (18) (2022) 7034, http://dx.doi.org/10.3390/s22187034.

[26] E. Yuda, M. Shibata, Y. Ogata, N. Ueda, T. Yambe, M. Yoshizawa, J. Hayano, Pulse rate variability: A new biomarker, not a surrogate for heart rate variability, J. Physiol. Anthropol. 39 (1) (2020) 21, http://dx.doi.org/10.1186/s40101-020-00233-x.

[27] J. Nikolic-Popovic, R. Goubran, Impact of motion artifacts on heart rate variability measurements and classification performance, in: 2013 IEEE International Symposium on Medical Measurements and Applications, MeMeA, IEEE, Gatineau, QC, 2013, pp. 156–159, http://dx.doi.org/10.1109/MeMeA.2013.6549726.

[28] C. Kirschbaum, K.M. Pirke, D. Hellhammer, The 'Trier Social Stress Test' – A tool for investigating psychobiological stress responses in a Laboratory Setting, Neuropsychobiology 28 (1–2) (1993) 76–81, http://dx.doi.org/10.1159/000119004.

[29] W. Lovallo, The cold pressor test and autonomic function: A review and integration, Psychophysiology 12 (3) (1975) 268–282, http://dx.doi.org/10.1111/j.1469-8986.1975.tb01289.x.

[30] S. Kumar, O. Tansley-Hancock, W. Sedley, J.S. Winston, M.F. Callaghan, M. Allen, T.E. Cope, P.E. Gander, D.E. Bamiou, T.D. Griffiths, The brain basis for misophonia, Curr. Biol. 27 (4) (2017) 527–533, http://dx.doi.org/10.1016/j.cub.2016.12.048.

[31] E. Grossini, A. Stecco, C. Gramaglia, D. De Zanet, R. Cantello, B. Gori, D. Negroni, D. Azzolina, D. Ferrante, A. Feggi, A. Carriero, P. Zeppegno, Misophonia: Analysis of the neuroanatomic patterns at the basis of psychiatric symptoms and changes of the orthosympathetic/ parasympathetic balance, Front. Neurosci. 16 (2022) 827998, http://dx.doi.org/10.3389/fnins.2022.827998.

[32] A. Ferrer-Torres, L. Giménez-Llort, Sounds of silence in times of COVID-19: Distress and loss of cardiac coherence in people with misophonia caused by real, imagined or evoked triggering sounds, Front. Psychiatry 12 (2021) 638949, http://dx.doi.org/10.3389/fpsyt.2021.638949.

[33] D. Shepherd, Electrophysiological approaches to noise sensitivity, J. Clin. Exp. Neuropsychol. (2016) 14, http://dx.doi.org/10.1080/13803395.2016.1176995.

[34] P. Reinhart, K. Griffin, C. Micheyl, Changes in heart rate variability following acoustic therapy in individuals with tinnitus, J. Speech, Lang. Hear. Res. (2021) 1–7, http://dx.doi.org/10.1044/2021_JSLHR-20-00596.

[35] B. Pfeiffer, L. Stein Duker, A. Murphy, C. Shui, Effectiveness of noise-attenuating headphones on physiological responses for children with autism spectrum disorders, Front. Integr. Neurosci. 13 (2019) 65, http://dx.doi.org/10.3389/fnint.2019.00065.

[36] M. Edelstein, D. Brang, R. Rouw, V.S. Ramachandran, Misophonia: Physiological investigations and case descriptions, Front. Hum. Neurosci. 7 (2013) http://dx.doi.org/10.3389/fnhum.2013.00296.

[37] E.J. Choi, Y. Yun, S. Yoo, K.S. Kim, J.S. Park, I. Choi, Autonomic conditions in tinnitus and implications for Korean medicine, Evidence-Based Complement. Altern. Med. 2013 (2013) 1–5, http://dx.doi.org/10.1155/2013/402585.

[38] J. Choi, B. Ahmed, R. Gutierrez-Osuna, Development and evaluation of an ambulatory stress monitor based on wearable sensors, IEEE Trans. Inf. Technol. Biomed. 16 (2) (2012) 279–286, http://dx.doi.org/10.1109/TITB.2011.2169804.

[39] V. Mishra, S. Sen, G. Chen, T. Hao, J. Rogers, C.H. Chen, D. Kotz, Evaluating the reproducibility of physiological stress detection models, Proc. ACM Interact., Mob., Wearable Ubiquitous Technol. 4 (4) (2020) 1–29, http://dx.doi.org/10.1145/3432220.

[40] A.O. Akmandor, N.K. Jha, Keep the stress away with SoDA: Stress detection and alleviation system, IEEE Trans. Multi-Scale Comput. Syst. 3 (4) (2017) 269–282, http://dx.doi.org/10.1109/TMSCS.2017.2703613.

[41] H.G. Kim, E.J. Cheon, D.S. Bai, Y.H. Lee, B.H. Koo, Stress and heart rate variability: A meta-analysis and review of the literature, Psychiatry Investig. 15 (3) (2018) 235–245, http://dx.doi.org/10.30773/pi.2017.08.17.

[42] F. Shaffer, J.P. Ginsberg, An overview of heart rate variability metrics and norms, Front. Public Health 5 (2017) 258, http://dx.doi.org/10.3389/fpubh.2017.00258.

[43] T. Pham, Z.J. Lau, S.H.A. Chen, D. Makowski, Heart rate variability in psychology: A review of HRV indices and an analysis tutorial, Sensors 21 (12) (2021) 3998, http://dx.doi.org/10.3390/s21123998.

[44] S. Laborde, E. Mosley, J.F. Thayer, Heart rate variability and cardiac vagal tone in psychophysiological research – recommendations for experiment planning, data analysis, and data reporting, Front. Psychol. 08 (2017) http://dx.doi.org/10.3389/fpsyg.2017.00213.

[45] K. Shi, T. Steigleder, S. Schellenberger, F. Michler, A. Malessa, F. Lurz, N. Rohleder, C. Ostgathe, R. Weigel, A. Koelpin, Contactless analysis of heart rate variability during cold pressor test using radar interferometry and bidirectional LSTM networks, Sci. Rep. 11 (1) (2021) 3025, http://dx.doi.org/10.1038/s41598-021-81101-1.

[46] S. Ollander, C. Godin, A. Campagne, S. Charbonnier, A comparison of wearable and stationary sensors for stress detection, in: 2016 IEEE International Conference on Systems, Man, and Cybernetics, SMC, IEEE, Budapest, Hungary, 2016, pp. 004362–004366, http://dx.doi.org/10.1109/SMC.2016.7844917.

[47] J. Voix, F. Laville, Expandable Earplug with Smart Custom Fitting Capabilities. Vol. 111, 2003. International Institute of Noise Control Engineering (I-INCE) Ames, IA, Dearborn, MI, USA., 2002, pp. pp. 833–841.

[48] R.C. Brindle, A.T. Ginty, A.C. Phillips, J.P. Fisher, D. McIntyre, D. Carroll, Heart rate complexity: A novel approach to assessing cardiac stress reactivity: Cardiac stress reactivity and heart rate complexity, Psychophysiology 53 (4) (2016) 465–472, http://dx.doi.org/10.1111/psyp.12576.

[49] Z. Visnovcova, M. Mestanik, M. Javorka, D. Mokra, M. Gala, A. Jurko, A. Calkovska, I. Tonhajzerova, Complexity and time asymmetry of heart rate variability are altered in acute mental stress, Physiol. Meas. 35 (7) (2014) 1319–1334, http://dx.doi.org/10.1088/0967-3334/35/7/1319.

[50] A. Monaco, R. Cattaneo, E. Ortu, M.V. Constantinescu, D. Pietropaoli, Sensory trigeminal ULF-TENS stimulation reduces HRV response to experimentally induced arithmetic stress: A randomized clinical trial, Physiol. Behav. 173 (2017) 209–215, http://dx.doi.org/10.1016/j.physbeh.2017.02.014.

[51] V. Mishra, G. Pope, S. Lord, S. Lewia, B. Lowens, K. Caine, S. Sen, R. Halter, D. Kotz, The case for a commodity hardware solution for stress detection, in: Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, ACM, Singapore, Singapore, 2018, pp. 1717–1728, http://dx.doi.org/10.1145/3267305.3267538.

[52] R.E. Bouserhal, T.H. Falk, J. Voix, In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension, J. Acoust. Soc. Am. 141 (3) (2017) 1321–1331, http://dx.doi.org/10.1121/1.4976051.

[53] A. Brugnera, C. Zarbo, M.P. Tarvainen, P. Marchettini, R. Adorni, A. Compare, Heart rate variability during acute psychosocial stress: A randomized cross-over trial of verbal and non-verbal laboratory stressors, Int. J. Psychophysiol. 127 (2018) 17–25, http://dx.doi.org/10.1016/j.ijpsycho.2018.02.016.

[54] K.P. Waye, J. Bengtsson, R. Rylander, F. Hucklebridge, P. Evans, A. Clow, Low frequency noise enhances cortisol among noise sensitive subjects during work performance, Life Sci. 70 (7) (2002) 745–758, http://dx.doi.org/10.1016/S0024-3205(01)01450-3.

[55] S. Hébert, S.J. Lupien, Salivary cortisol levels, subjective stress, and tinnitus intensity in tinnitus sufferers during noise exposure in the laboratory, Int. J. Hygiene Environ. Health 212 (1) (2009) 37–44, http://dx.doi.org/10.1016/j.ijheh.2007.11.005.

[56] D.R. Bach, H. Schachinger, J.G. Neuhoff, F. Esposito, F.D. Salle, C. Lehmann, M. Herdener, K. Scheffler, E. Seifritz, Rising sound intensity: An intrinsic warning cue activating the amygdala, Cerebral Cortex 18 (1) (2008) 145–150, http://dx.doi.org/10.1093/cercor/bhm040.

[57] M.G. Pagé, L. Dassieu, E. Develay, M. Roy, E. Vachon-Presseau, S. Lupien, P. Rainville, The stressful characteristics of pain that drive you NUTS: A qualitative exploration of a stress model to understand the chronic pain experience, Pain Med. 22 (5) (2021) 1095–1108, http://dx.doi.org/10.1093/pm/pnaa370.

[58] A. Baigi, A. Oden, V. Almlid-Larsen, M.-L. Barrenäs, K.M. Holgers, Tinnitus in the general population with a focus on noise and stress: A public health study, Ear Hear. 32 (6) (2011) 787–789, http://dx.doi.org/10.1097/AUD.0b013e31822229bd.

[59] E.M. Colagrosso, P. Fournier, E.M. Fitzpatrick, S. Hébert, A qualitative study on factors modulating tinnitus experience, Ear Hear. 40 (3) (2019) 636–644, http://dx.doi.org/10.1097/AUD.0000000000000642.

[60] L. Mourot, M. Bouhaddi, J. Regnard, Effects of the cold pressor test on cardiac autonomic control in normal subjects, Physiol. Res. (2009) 83–91, http://dx.doi.org/10.33549/physiolres.931360.

[61] M. Malik, J.T. Bigger, A.J. Camm, R.E. Kleiger, A. Malliani, A.J. Moss, P.J. Schwartz, Heart rate variability: Standards of measurement, physiological interpretation, and clinical use, Eur. Heart J. 17 (3) (1996) 354–381, http://dx.doi.org/10.1093/oxfordjournals.eurheartj.a014868.

[62] N. Momeni, A. Arza, J. Rodrigues, C. Sandi, D. Atienza, CAFS: Cost-aware features selection method for multimodal stress monitoring on wearable devices, IEEE Trans. Biomed. Eng. (2021) http://dx.doi.org/10.1109/TBME.2021.3113593, 1–1.

[63] K. Hovsepian, M. al'Absi, E. Ertin, T. Kamarck, M. Nakajima, S. Kumar, 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 - UbiComp '15, ACM Press, Osaka, Japan, 2015, pp. 493–504, http://dx.doi.org/10.1145/2750858.2807526.

[64] K. Suzuki, T. Laohakangvalvit, R. Matsubara, M. Sugaya, Constructing an emotion estimation model based on EEG/HRV Indexes using feature extraction and feature selection algorithms, Sensors 21 (9) (2021) 2910, http://dx.doi.org/10.3390/s21092910.

[65] F.T. Sun, C. Kuo, H.T. Cheng, S. Buthpitiya, P. Collins, M.L. Griss, Activity-aware mental stress detection using physiological sensors, in: International Conference on Mobile Computing, Applications, and Services, 2012, p. 20, http://dx.doi.org/10.1007/978-3-642-29336-8_12.

[66] M.M. Pulopulos, M.A. Vanderhasselt, R. De Raedt, Association between changes in heart rate variability during the anticipation of a stressful situation and the stress-induced cortisol response, Psychoneuroendocrinology 94 (2018) 63–71, http://dx.doi.org/10.1016/j.psyneuen.2018.05.004.

[67] D. Cajal, D. Hernando, J. Lázaro, P. Laguna, E. Gil, R. Bailón, Effects of missing data on heart rate variability metrics, Sensors 22 (15) (2022) 5774, http://dx.doi.org/10.3390/s22155774.

[68] D. Makowski, T. Pham, Z.J. Lau, J.C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, S.H.A. Chen, NeuroKit2: A Python toolbox for neurophysiological signal processing, Behav. Res. Methods 53 (4) (2021) 1689–1696, http://dx.doi.org/10.3758/s13428-020-01516-y.

[69] P.M. McEvoy, M.P. Hyett, A.R. Johnson, D.M. Erceg-Hurn, P.J. Clarke, M.J. Kyron, S.R. Bank, L. Haseler, L.M. Saulsman, M.L. Moulds, J.R. Grisham, E.A. Holmes, D.A. Moscovitch, O.V. Lipp, R.M. Rapee, Impacts of imagery-enhanced versus verbally-based cognitive behavioral group therapy on psychophysiological parameters in social anxiety disorder: Results from a randomized-controlled trial, Behav. Res. Ther. 155 (2022) 104131, http://dx.doi.org/10.1016/j.brat.2022.104131.

[70] P. Chabot, R.E. Bouserhal, P. Cardinal, J. Voix, Detection and classification of human-produced nonverbal audio events, Appl. Acoust. 171 (2021) 107643, http://dx.doi.org/10.1016/j.apacoust.2020.107643.

[71] M. Parent, A. Tiwari, I. Albuquerque, J.-F. Gagnon, D. Lafond, S. Tremblay, T.H. Falk, A multimodal approach to improve the robustness of physiological stress prediction during physical activity, in: 2019 IEEE International Conference on Systems, Man and Cybernetics, SMC, IEEE, Bari, Italy, 2019, pp. 4131–4136, http://dx.doi.org/10.1109/SMC.2019.8914254.

[72] M. Nardelli, G. Valenza, A. Greco, A. Lanata, E.P. Scilingo, Recognizing emotions induced by affective sounds through heart rate variability, IEEE Trans. Affect. Comput. 6 (4) (2015) 385–394, http://dx.doi.org/10.1109/TAFFC.2015.2432810.

[73] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. (2011) 6.

[74] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco, California, USA, 2016, pp. 785–794, http://dx.doi.org/10.1145/2939672.2939785.

[75] M. Sadeghi, A.D. McDonald, F. Sasangohar, Posttraumatic stress disorder hyperarousal event detection using smartwatch physiological and activity data, in: M.S. Kaiser (Ed.), PLoS One 17 (5) (2022) e0267749, http://dx.doi.org/10.1371/journal.pone.0267749.

[76] A.J. Masino, D. Forsyth, H. Nuske, J. Herrington, J. Pennington, Y. Kushleyeva, C.P. Bonafide, M-health and autism: Recognizing stress and anxiety with machine learning and wearables data, in: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems, CBMS, IEEE, Cordoba, Spain, 2019, pp. 714–719, http://dx.doi.org/10.1109/CBMS.2019.00144.

[77] A. Gupta, J. Jain, S. Poundrik, M.K. Shetty, M.P. Girish, M.D. Gupta, Interpretable AI model-based predictions of ECG changes in COVID-recovered patients, in: 2021 4th International Conference on Bio-Engineering for Smart Technologies, BioSMART, IEEE, Paris / Créteil, France, 2021, pp. 1–5, http://dx.doi.org/10.1109/BioSMART54244.2021.9677747.

[78] C. Molnar, Interpretable Machine Learning, Lulu. com, 2020.

[79] W.S. Liew, C.K. Loo, S. Wermter, Emotion recognition using explainable genetically optimized fuzzy ART ensembles, IEEE Access 9 (2021) 61513–61531, http://dx.doi.org/10.1109/ACCESS.2021.3072120.

[80] J. Guo, Y. Dai, C. Wang, H. Wu, T. Xu, K. Lin, A physiological data-driven model for learners' cognitive load detection using HRV-PRV feature fusion and optimized XGBoost classification, Softw. - Pract. Exp. 50 (11) (2020) 2046–2064, http://dx.doi.org/10.1002/spe.2730.

[81] F. Poursabzi-Sangdeh, D.G. Goldstein, J.M. Hofman, J.W. Wortman Vaughan, H. Wallach, Manipulating and measuring model interpretability, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, ACM, Yokohama, Japan, 2021, pp. 1–52, http://dx.doi.org/10.1145/3411764.3445315.

[82] C.J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key challenges for delivering clinical impact with artificial intelligence, BMC Med. 17 (1) (2019) 195, http://dx.doi.org/10.1186/s12916-019-1426-2.

[83] T. Pham, Z.J. Lau, S.A. Chen, D. Makowski, Unveiling the Structure of Heart Rate Variability (HRV) Indices: A Data-driven Meta-clustering Approach, 2021, http://dx.doi.org/10.31234/osf.io/mwa6x, Preprint. PsyArXiv.

[84] M. Hosseini, M. Powell, J. Collins, C. Callahan-Flintoft, W. Jones, H. Bowman, B. Wyble, I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data, Neurosci. Biobehav. Rev. 119 (2020) 456–467, http://dx.doi.org/10.1016/j.neubiorev.2020.09.036.

[85] J.R. Zech, M.A. Badgeley, M. Liu, A.B. Costa, J.J. Titano, E.K. Oermann, Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study, in: A. Sheikh (Ed.), PLOS Med. 15 (11) (2018) e1002683, http://dx.doi.org/10.1371/journal.pmed.1002683.

[86] C.H. Chang, G.A. Adam, A. Goldenberg, Towards robust classification model by counterfactual and invariant data generation, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, Nashville, TN, USA, 2021, pp. 15207–15216, http://dx.doi.org/10.1109/CVPR46437.2021.01496.

[87] V. Agarwal, R. Shetty, M. Fritz, Towards causal VQA: Revealing and reducing spurious correlations by invariant and covariant semantic editing, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, Seattle, WA, USA, 2020, pp. 9687–9695, http://dx.doi.org/10.1109/CVPR42600.2020.00971.

[88] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Gender bias in coreference resolution: Evaluation and debiasing methods, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 15–20, http://dx.doi.org/10.18653/v1/N18-2003.

[89] H. Yu, A. Sano, Semi-supervised learning and data augmentation in wearable-based momentary stress detection in the wild, 2022, arXiv:2202.12935.

[90] G. Bernal, S.M. Montgomery, P. Maes, Brain-computer interfaces, open-source, and democratizing the future of augmented consciousness, Front. Comput. Sci. 3 (2021) 661300, http://dx.doi.org/10.3389/fcomp.2021.661300.