UNIVERSITÄT AUGSBURG
Universitätsbibliothek

OPEN ACCESS
OPUS AUGSBURG

Stressor type matters! - Exploring factors influencing cross-dataset generalizability of physiological stress detection

Pooja Prajod, Bhargavi Mahesh, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Prajod, Pooja, Bhargavi Mahesh, and Elisabeth André. 2024. "Stressor type matters! - Exploring factors influencing cross-dataset generalizability of physiological stress detection." In *ICMI '24: proceedings of the 26th International Conference on Multimodal Interaction, San Jose, Costa Rica, 4-8 November 2024*, 508–17. New York, NY: ACM. https://doi.org/10.1145/3678957.3685738.

Nutzungsbedingungen / Terms of use:

CC BY 4.0

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
CC-BY 4.0: Creative Commons: Namensnennung
Weitere Informationen finden Sie unter: / For more information see:
https://creativecommons.org/licenses/by/4.0/deed.de

Stressor Type Matters! — Exploring Factors Influencing Cross-Dataset Generalizability of Physiological Stress Detection

Pooja Prajod
University of Augsburg
Augsburg, Germany
pooja.prajod@uni-a.de

Bhargavi Mahesh
University of Augsburg
Augsburg, Germany
bhargavi.mahesh@uni-a.de

Elisabeth André
University of Augsburg
Augsburg, Germany
elisabeth.andre@uni-a.de

Abstract

Automatic stress detection using heart rate variability (HRV) features has gained significant traction as it utilizes unobtrusive wearable sensors measuring signals like electrocardiogram (ECG) or blood volume pulse (BVP). However, detecting stress through such physiological signals presents a considerable challenge owing to the variations in recorded signals influenced by factors, such as perceived stress intensity and measurement devices. Consequently, stress detection models developed on one dataset may perform poorly on unseen data collected under different conditions. To address this challenge, this study explores the generalizability of machine learning models trained on HRV features for binary stress detection. Our goal extends beyond evaluating generalization performance; we aim to identify the characteristics of datasets that have the most significant influence on generalizability. We leverage four publicly available stress datasets (WESAD, SWELL-KW, ForDigitStress, VerBIO) that vary in at least one of the characteristics such as stress elicitation techniques, stress intensity, and sensor devices. Employing a cross-dataset evaluation approach, we explore which of these characteristics strongly influence model generalizability. Our findings reveal a crucial factor affecting model generalizability: primary stressor. Models achieved good performance across datasets when the primary stressor (e.g., social evaluation in our case) remains consistent. Factors like stress intensity or brand of the measurement device had minimal impact on cross-dataset performance. Based on our findings, we recommend matching the primary stressor when deploying HRV-based stress models in new environments. Although previous works have performed cross-dataset evaluation of stress models, this is the first study to systematically investigate the factors influencing the cross-dataset applicability of HRV-based stress models. Our insights are crucial for scenarios with limited data, where techniques like domain generalization and domain adaptation may not be applicable.

CCS Concepts

• **Computing methodologies** → **Cross-validation**; *Machine learning*; • **Human-centered computing** → *Empirical studies in ubiquitous and mobile computing*.

This work is licensed under a Creative Commons Attribution International 4.0 License.

ICMI '24, November 04–08, 2024, San Jose, Costa Rica
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3685738

Keywords

Stress, Generalizability, Cross-dataset, Machine learning, Heart rate variability, Electrocardiography, Photoplethysmography

ACM Reference Format:
Pooja Prajod, Bhargavi Mahesh, and Elisabeth André. 2024. Stressor Type Matters! — Exploring Factors Influencing Cross-Dataset Generalizability of Physiological Stress Detection. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 04–08, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3678957.3685738

1 Introduction

The ability to detect stress in real-time has become increasingly important within the fields of affective computing and human-machine interaction (HMI) [3]. Early stress detection offers a valuable tool for promoting well-being and potentially preventing long-term health consequences [1, 3, 16]. Consequently, a growing area of research focuses on developing interactive systems that can not only detect stress but also provide personalized interventions to manage it effectively [6, 44].

Researchers have explored various modalities for stress detection, including psychological tools, behavioral patterns, and physiological signals [3, 15]. However, physiological signals are considered more reliable than the other methods as they eliminate certain measurement biases. Moreover, the growing popularity of unobtrusive wearable sensors further facilitates continuous stress monitoring through physiological signals.

Stress manifests differently depending on the context [3]. For instance, the stress experienced during an exam likely differs from that of public speaking. While controlled lab settings are often used to develop stress detection models, real-world HMI scenarios are far more diverse. Therefore, assessing the generalizability of these models - their ability to perform well in unseen contexts - is crucial for real-world applications. Cross-dataset evaluation, where a model trained on one dataset is tested on another, is a common approach to assess generalizability. Good cross-dataset performance suggests broader applicability of the model. This method of assessing generalizability has been employed in other contexts such as mood prediction [26], pain detection [33], and activity recognition [4, 34].

The existing literature on stress detection models acknowledges the importance of generalizability, with a few studies exploring cross-dataset evaluations [41]. These studies typically investigate the applicability of models across different datasets and sometimes explore combining datasets for improved performance (refer to Section 2 for details). While prior works often report limited cross-dataset performance, a key gap exists: a lack of systematic investigation of the factors influencing this limited generalizability.

We address this research gap by conducting extensive cross-dataset evaluations using multiple stress datasets. We train three popular machine learning models - random forest classifier (RFC), support vector machine (SVM), and multi-layer perceptron (MLP) - using heart rate variability (HRV) features. We aim to identify the characteristics of these datasets that significantly impact model generalizability. Our findings are crucial for developing stress detection models with broader applicability in HMI systems.

Our study is particularly important for scenarios where researchers aim to improve the interaction using existing stress datasets. For example, in real-world industrial HRI settings, collecting stress data without disrupting their current setup is challenging [28]. However, it is feasible to get information about stressful scenarios through worker interviews. Our findings contribute to a guideline on what characteristics to consider when choosing an existing stress dataset.

2 Related Work

The publicly available datasets such as WESAD [38] and SWELL-KW [21], led to the development of numerous stress detection models [12, 18]. Some of these studies (e.g., [10]) focused on comparing multiple models to determine the best model for stress detection. However, many of these comparison studies are trained and evaluated on the same datasets. In this section, we discuss some of the existing works that performed cross-dataset evaluations on stress detection models.

Mishra et al. [27] conducted cross-dataset evaluations across four cognitive stress datasets - two had electrocardiogram (ECG) signals, and the other two had blood volume pulse (BVP) signals. In addition to arithmetic tasks, the ECG datasets had a startle response test and cold pressor task as stressors. The authors trained SVM models using HRV features extracted from the mental stress segments of these datasets. While the ECG-based HRV models performed well in detecting stress in each other's arithmetic tasks, they had a performance drop of 15 - 30% in predicting stress in the BVP datasets. The authors attributed this drop in performance, despite having the same stressor, to the difference in sensors. They also noticed an approximately 20 - 40% drop in the performance when detecting overall stress (including startle and cold pressor segments), even within the same datasets. Their findings suggest that the models trained on one type of stressor may not be efficient in detecting other stress responses.

Prajod and André [32] trained ECG-based deep-learning models and HRV-based shallow models (RFC, SVM, MLP) on WESAD and SWELL-KW datasets. While deep learning models outperformed other methods in within-dataset evaluations, they performed poorly in cross-dataset evaluations (WESAD models tested on SWELL-KW and vice versa). Although the HRV-based models performed better than deep learning models in cross-dataset evaluations, their performances were still low. Moreover, the combining datasets did not improve the model's performance on individual datasets. Similarly, Albaladejo-González et al. [2] trained HRV-based models on the WESAD and SWELL-KW datasets. They also observed poor cross-dataset performances and a lack of stress detection improvements by combining datasets.

Benchekroun et al. [8] trained two HRV-based models (RFC, logistic regression) on the MMSD [7] and UWS [39] datasets. They tested the MMSD models using the UWS data and found that the f1-scores were 12 – 14% lower than the UWS models (from within-dataset evaluations). They further noted that the f1-score for stress class was very low (less than 50 %), meaning the models were not very efficient in detecting stressful instances.

Vos et al. [40] trained shallow models (RFC, SVM, XGBoost) using heart rate and electro-dermal activity (EDA) features from the SWELL-KW dataset and evaluated them on WESAD and NEURO [9] datasets. All three models showed poor cross-dataset performances. They also implemented an ensemble model and repeated the cross-dataset evaluation. Although this ensemble model yielded a slight improvement, the performance was still poor (f1-score $< 50\%$). Furthermore, they trained the ensemble model using a combined dataset (SWELL-KW, NEURO, UBFC-Phys [36]) and evaluated it on WESAD. While the accuracy increased slightly, the f1-score dropped further. Their experiments highlight the challenges of developing a generic stress model.

The above works primarily focused on assessing the generalizability of HRV-based models. However, Liapis et al. [22] demonstrated that EDA-based shallow models also struggle with generalizability. They trained their models on the WESAD dataset and then evaluated them on their own dataset. Notably, their dataset contained subtle stress instances, unlike WESAD, implying that the stress intensity might further impact generalizability.

Baird et al. [5] compared models trained on speech features from three social stress datasets (FAU-TSST, Ulm-TSST, and Reg-TSST). These datasets induced stress using the Trier social stress test (TSST). They predicted cortisol levels as a proxy for stress levels. In cross-dataset evaluations, the trends of predicted cortisol levels were aligned for the models, indicating compatibility between datasets. Due to the dataset compatibility, they suggested that training models using data from both these datasets could result in better-performing models.

Table 1 presents an overview of the existing studies on the cross-dataset generalizability of stress detection models. Most studies assess the generalizability of a model to determine if it is applicable in other stress scenarios. Some studies take a step further to evaluate whether combining datasets to train models improves stress detection performances. Although most studies observe low generalizability of stress models, few studies provide insights into plausible factors influencing the models' performance. For example, Mishra et al. highlighted the poor performance of mental stress models in detecting physical stress. Similarly, Liapis et al. noted the difference in stress intensities of the evaluated datasets. Prajod and André hinted at multiple factors such as stress intensity and measurement devices that could affect generalizability. However, these studies did not further investigate these factors. Hence, a crucial question remains relatively unexplored - What factors or characteristics of the stress datasets need to match for cross-dataset applicability of models?

3 Materials and Methods

3.1 Datasets

In this study, we focus on binary stress detection (stress vs. no-stress) using HRV features. We utilize HRV as it is one of the popular modalities in real-time stress recognition. We leverage four

Table 1: An overview of the existing works that perform cross-dataset evaluations of their stress models

Paper	Input	Datasets	Aim
Mishra et al. [27]	ECG, BVP	Own datasets (mainly mental arithmetic tasks)	Assess generalizability
Liapis et al. [22]	EDA	WESAD, Own UX stress dataset	Assess generalizability
Baird et al. [5]	Speech	Own TSST datasets	Assess compatibility for combining datasets
Prajod and André [32]	ECG	WESAD, SWELL-KW	Assess generalizability, Combine datasets
Albaladejo-González et al. [2]	ECG	WESAD, SWELL-KW	Assess generalizability, Combine datasets
Benchekroun et al. [8]	ECG	MMSD, UWS	Assess generalizability
Vos et al. [40]	Heart rate, EDA	WESAD, SWELL-KW, NEURO, UBFC-Phys	Assess generalizability, Combine datasets
This work	ECG, BVP	WESAD, SWELL-KW, ForDigitStress, VerBIO	Assess generalizability, Combine datasets, **Identify factors influencing generalizability**

publicly available stress datasets: WESAD [38], SWELL-KW [21], ForDigitStress [19], and VerBIO [43]. While SWELL-KW contains ECG signals that can be used for extracting HRV, the ForDigitStress dataset has BVP signals for extracting HRV. The WESAD and VerBIO datasets contain both ECG and BVP signals.

While we tried to perform an exhaustive comparison, we excluded some comparisons for brevity. This choice was based on whether an additional comparison investigates a unique factor. SWELL-WESAD did not have many similarities; WESAD-ForDigit were similar in stressor; ForDigit-VerBIO were similar in stressor and intensity; WESAD-VerBIO were similar in stressor, elicitation, and sensor. SWELL-VerBIO is similar to the SWELL-WESAD comparison, differing in most considered factors. However, we consider only the BVP signals from the VerBIO dataset to remove some redundant comparisons in our analysis.

We note that EDA is another common modality, which is part of all four datasets. However, EDA data resulted in a low baseline performance in some of the datasets (e.g., ForDigitStress)), which could potentially indicate noisy signals due to arm movements. A low baseline performance hinders the assessment of generalizability as any drop in performance could be due to signal quality.

3.1.1 ***WESAD.*** This is a multimodal stress and affect dataset containing various physiological signals, including ECG, EDA, and BVP. The data from 15 participants were collected using a chest-worn RespiBan and a wrist-worn Empatica E4 device. This investigation utilizes the ECG data recorded by the chest-worn device at 700 Hz.

The participants were subject to three conditions: neutral, amusement, and stress. In the stress condition, the participants experienced social stress induced by the TSST technique. The participants engaged in public speaking and mental arithmetic tasks while being evaluated by a three-member panel. To induce amusement, the participants watched selected funny video clips. The experimental sessions began with the neutral condition, followed by the stress and amusement conditions in alternating order. For each participant, the neutral condition lasted for approximately 20 minutes, the stress condition for 10 minutes, and the amusement condition for around 6.5 minutes.

We focus on stress detection, i.e., distinguishing between stress and no-stress samples. Following the labeling scheme proposed by the dataset creators, data from both neutral and amusement conditions were considered as no-stress samples.

3.1.2 ***SWELL-KW.*** This is also a multimodal stress dataset that contains two physiological signals, ECG and EDA. This dataset consists of data from 25 participants who engaged in typical knowledge tasks like writing reports and presentations. The ECG data was collected using the TMSI Mobi device at a sampling rate of 2048 Hz.

The participants underwent three experimental conditions: neutral, email interruptions, and time pressure. During the email interruption session, participants received eight emails, many irrelevant and some requiring responses. In the time pressure condition, participants had to complete the tasks within two-thirds of the allotted neutral session time. Like the WESAD dataset, the first session was always neutral, followed by the other two conditions in alternating order. The neutral and email interruption sessions lasted approximately 45 minutes, while the time pressure session was around 30 minutes long.

Notably, the participants did not report experiencing high stress in any of the three conditions. However, they indicated a higher temporal demand during the time pressure session. While training stress detection models, the dataset creators considered the data from email interruptions and time pressure sessions as stress samples and the neutral session as no-stress samples [20]. Therefore, we follow the same labeling scheme for consistency. However, three participants were excluded due to missing data.

3.1.3 ***ForDigitStress.*** This dataset represents another multimodal stress dataset with various behavioral (facial expression, body pose, etc.) and physiological (BVP, EDA) signals. The dataset was collected from 40 participants who attended a mock job interview session. The BVP data was collected using an IOM-biofeedback device, operating at 27 Hz.

The experimental session was divided into three phases: preparation, interview, and post-interview. The participants were asked to submit their resumes in advance so that the interviewer could customize the questions depending on the participant. During the interview phase, the interviewer questioned the participants on topics such as their strengths/weaknesses, salary expectations, and job-related scenarios. The interview phase lasted for about 25 minutes. In addition to self-reported stress levels, the participants' saliva samples were collected to assess their cortisol levels. The cortisol levels served as ground truths for the presence of stress.

Although both the preparation and post-interview phases can be considered as no-stress conditions, the dataset creators suggest using data from later parts of the post-interview phase. So, we utilize the interview phase as stress samples and the last segments of the post-interview phase (15 - 20 mins) as no-stress samples.

3.1.4 ***VerBIO.*** This dataset was collected to investigate if exposing participants to public speaking through virtual reality (VR) training would reduce public speaking anxiety. The dataset contains data from 55 participants who were recruited for two real and eight virtual oral presentations over two days. The physiological data from two wearable sensors, Empatica E4 (BVP, EDA, skin temperature) and Actiwave Cardio Monitor (ECG) was acquired. Several self-assessment questionnaires were employed to capture demographics, state- and trait-based psychological measures.

The experiment was divided into three sessions: pre-interview (Pre), interview (Test), and post-interview (Post) sessions. The Pre and Post sessions involved presenting in front of a real audience, whereas the Test session involved a VR audience. Each session further constituted relaxation, preparation, and presentation phases. The presentation topics were of general interest and were about four minutes long.

We note that the participation dropped through the three sessions, with maximum participation in the "Pre" session and the lowest in the "Post" session. To avoid drastic imbalances in data between participants, we utilized only the data from the "Pre" session. We considered the data belonging to the relaxation phases as no-stress and those of the presentation phase as stressful. In our analysis, we normalize the data for each participant using a few minutes of no-stress data (see Section 3.3). Hence, 10 participants with a low amount of no-stress data were excluded.

3.1.5 ***Comparing Dataset Characteristics.*** A brief overview of the four datasets is presented in Table 2. Beyond sensor differences, key variations exist in participant numbers and the duration of stress/no-stress samples. The SWELL-KW and ForDigitStress datasets are considerably larger than the VerBIO dataset. Stress intensity levels also vary across the datasets. WESAD reports high-stress experiences, while ForDigitStress and VerBIO involve moderate stress, and SWELL-KW reports low stress.

Another differentiating factor is the stress elicitation technique. While WESAD used a standardized test (TSST), the other three datasets leveraged simulations of natural scenarios. However, all datasets were collected in controlled laboratory environments. Based on the description of the stressor provided by the dataset creators, we also highlight the primary stressor in each dataset. While tasks like TSST and public speaking involve some cognitive effort, the primary stressor remains social evaluation in these cases.

Among the social stress datasets, the WESAD and VerBIO datasets share more similarities than the ForDigitStress dataset. Both datasets utilized the Empatica E4 device to measure the BVP signals. While WESAD employed TSST to elicit stress, VerBIO relied on a public speaking task that is a sub-task of the TSST protocol. On the other hand, the ForDigitStress dataset utilized a different measurement device and employed a mock job interview to induce stress.

3.2 Approach

Our investigations involve the three assessments listed below. The idea is to run a series of these assessments using the datasets described in Section 3.1 to investigate the dataset characteristics that considerably influence model generalizability.

(1) **Within-dataset Assessment:** involves training and evaluating models on the same dataset. We utilize the leave-one-subject-out (LOSO) technique to evaluate the models' performance on unseen participants of the same dataset.
(2) **Cross-dataset Assessment:** involves evaluating the models trained on one dataset using another dataset. This evaluation assesses to what extent these models can detect stress in new participants in different settings.
(3) **Combining Datasets:** involves training new models on a combined dataset consisting of data from two or more stress datasets, again using the LOSO technique. This step investigates potential improvements in models' performances due to an increase in the number and variations of the training data.

While ECG-based and BVP-based HRVs reflect similar physiological information related to heart activity, the models trained on these features might not perform equally well [17]. This discrepancy can be attributed to the BVP signals being more prone to noise from body movements than ECG [25]. This noise can affect the signal quality, which can, in turn, impact the stress detection performance. To avoid such influences in cross-dataset assessment, the above assessments are followed for ECG-based and BVP-based HRV models separately.

3.3 Data Processing

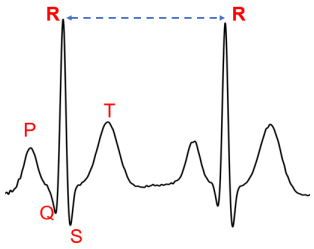

Figure 1: An example plot of ECG signal, marked with various repeated components of the signal.

3.3.1 ***ECG signals.*** The ECG signals are typically sampled at high frequencies and contain noises such as baseline wander (low-frequency, 0.5 - 0.6 Hz) and powerline interference (50 or 60 Hz). Figure 1 illustrates two beats from an ECG signal. Computing HRV

Table 2: An overview of some key characteristics of the four stress datasets

	WESAD	SWELL-KW	ForDigitStress	VerBIO
Stress elicitation technique	TSST	Interruptions, Time pressure	Job interview	Public speaking
Primary stressor	Social evaluation	Mental effort	Social evaluation	Social evaluation
Sensor	ECG: RespiBan, 700 Hz, BVP: Empatica E4, 64 Hz	ECG: TMSI Mobi, 2048 Hz	BVP: IOM, 27 Hz	BVP: Empatica E4, 64 Hz
Avg. stress level	18.5/24 (STAI index)	3.5/10 (Likert scale)	5.4/10 (Likert scale), 6.5/10 (Cortisol)	46.25/80 (CAI index)
Participants	15	22	40	45
Data duration (per participant)	stress: 10 mins, no-stress: 26.5 mins	stress: 75 mins, no-stress: 45 mins	stress: 25 mins, no-stress: 15 - 20 mins	stress: 2 - 5 mins, no-stress: 4 - 6 mins

relies on detecting the R peaks in the QRS complex of the ECG signal. We applied a second-order Butterworth band-pass filter with a frequency band of 8 - 20 Hz for optimal QRS signal-to-noise ratio [13].

For R-peak detection, we utilized the algorithm proposed by [13]. This algorithm is based on two key assumptions for healthy adults:

(a) A QRS complex contains one and only one heartbeat
(b) The duration of a typical QRS complex is in the range of 80 - 120 milliseconds

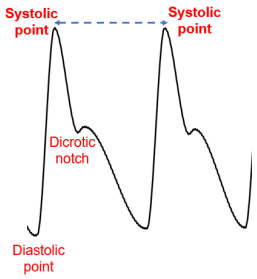

Figure 2: An example plot of BVP signal, marked with various repeated components of the signal.

3.3.2 ***BVP signals.*** Like ECG, the BVP signal is susceptible to baseline wander and high-frequency noise. Hence, a band-pass filter (0.5 - 8 Hz) was applied to reduce these noises [14]. Figure 2 illustrates two beats from a BVP signal. To derive the HRV signal from the BVP, the systolic peaks had to be detected. For this purpose, we employed the peak-finding algorithm from [19] to detect points that meet the following criteria:

(a) Amplitude threshold: The peaks had to be taller than a certain threshold. This threshold was set based on the distribution of peak heights in the entire signal.
(b) Distance between peaks: To avoid identifying every fluctuation as a peak, consecutive peaks had to be separated by a minimum interval of 0.333 seconds. This value corresponds to a maximum heart rate of 3 beats/second (180 beats/min).

3.3.3 ***HRV Features.*** Once the R-peaks (for ECG) or systolic peaks (for BVP) were identified, the time intervals between successive peaks were calculated to form the HRV signals. The features were calculated using 60-second segments with 59 seconds of overlap between consecutive segments.

A total of 22 well-known features [19, 30, 35, 38] were computed from the extracted HRV signals. These features belonged to the time domain (13 features), frequency domain (5 features), and poincaré plot characteristics (4 features). These features were calculated using the NeuroKit2 Python library [24].

The sensors used in the different datasets differ, potentially resulting in values recorded on different scales. In addition, the range of physiological recordings may vary from participant to participant [11, 29, 37]. To mitigate the effect of these differences, participant-specific Min-Max normalization was applied to each HRV feature. For real-time stress detection, the entire dataset would not be available for normalization. Similar to [23, 32], we used 5 minutes of neutral data to compute normalization parameters (minimum and maximum values) for each participant. For the VerBIO dataset, we used neutral data for normalization because many of the participants had less than 5 minutes of neutral data.

3.4 Machine Learning Models

We trained the following three machine learning models using the extracted HRV features, following the LOSO procedure. These three models were selected because they have been widely used in previous stress studies, including the original dataset papers. To account for the imbalanced sample distribution of the datasets, the "class_weight" hyperparameter for all models was set inversely proportional to the sample frequencies.

3.4.1 ***Random Forest Classifier (RFC).*** This is an ensemble learning method that combines predictions from multiple decision trees for improved performance and reduced overfitting. Each tree is trained on a subset of the available training set. The final prediction is determined by aggregating the predictions from all the trees (e.g., majority vote). This strategy often results in a better performance, even if the individual decision trees are weak predictors. A total of 200 decision trees (also called estimators) were trained, with a maximum depth of 5 for each tree.

3.4.2 ***Support Vector Machine (SVM).*** This is a commonly used supervised learning method for binary classification tasks. During the training process of this model, the objective is to find a hyperplane within the feature space that separates the data points

belonging to different classes. We utilized a linear SVM classifier with tolerance parameter 'C' set to 0.001.

3.4.3 ***Multi-layer Perceptron (MLP).*** This is a simple feed-forward neural network (also called simple artificial neural network), which has been growing in popularity for stress detection [2, 10, 45]. Our implementation followed an architecture consisting of an input layer, two hidden layers, and a prediction layer. A dropout layer (rate = 0.2) was included after the input layer to mitigate overfitting of the model. The two hidden layers (ReLU activation) followed the dropout layer, with 12 nodes for the first hidden layer and 6 nodes for the second hidden layer. This final layer outputs the classification result using a Sigmoid activation function.

This model was trained using the SGD optimizer (learning rate = 0.001) and weighted loss. It was also trained in batches of 256 samples. We utilized the early stopping technique, where the training stopped if the validation loss did not decrease for 15 consecutive iterations.

4 Results

We employed three evaluation strategies to assess model performances: within-dataset, cross-dataset, and combined dataset evaluations. For all evaluations, accuracy and f1-score metrics were used to quantify model performance. The detailed results for each evaluation strategy are presented below.

4.1 Within-dataset Assessment

We employed the LOSO technique to train and evaluate model performance within each of the four stress datasets. The average f1-score and accuracy for each dataset are presented in Table 3.

The VerBIO dataset yielded the best overall performance, with models achieving average accuracies and f1-scores exceeding 0.90. On the other hand, models trained on the SWELL-KW dataset exhibited lower average performance, with f1-scores and accuracies ranging from 0.60 to 0.70. All other datasets achieved average f1-scores above 0.75 and average accuracies greater than 0.80.

Except for VerBIO, MLP consistently achieved the highest performance within each dataset, followed by RFC and then SVM. For VerBIO, RFC outperformed all other models. Notably, all our models surpassed the baselines established in the respective dataset papers, except for the WESAD BVP-HRV models. The performance differences between models were relatively small, typically within a 0.05 margin.

Within the WESAD dataset, models trained on ECG-derived HRV features outperformed those using BVP-HRV features ($\text{pvalue}_{f1} > 0.1$, $\text{pvalue}_{acc} > 0.1$). However, the WESAD dataset paper [38] reported slightly higher performance for BVP-HRV models.

4.2 Cross-dataset Assessment

To assess the generalizability of the models, we conducted cross-dataset evaluations. Each LOSO model trained on one dataset was tested on unseen data from another dataset. For ECG-derived HRV features, we tested the SWELL-KW models on data from the WESAD dataset, and vice versa. The results for these cross-dataset evaluations are presented in Table 4 (SWELL-KW) and Table 5 (WESAD, ECG).

Table 3: Results of within-dataset LOSO evaluation of HRV models conducted on the four stress datasets

Model	F1-score	Accuracy
Test on SWELL-KW, ECG		
SWELL-KW baseline [20]	-	0.589
SWELL-KW RFC	0.634±0.151	0.664±0.139
SWELL-KW SVM	0.597±0.119	0.626±0.113
SWELL-KW MLP	**0.667**±0.137	**0.692**±0.130
Test on WESAD, ECG		
WESAD (ECG-HRV) baseline [38]	0.813	0.854
WESAD (ECG-HRV) RFC	0.819±0.161	0.863±0.110
WESAD (ECG-HRV) SVM	0.814±0.154	0.862±0.103
WESAD (ECG-HRV) MLP	**0.844**±0.169	**0.880**±0.126
Test on WESAD, BVP		
WESAD (BVP-HRV) baseline [38]	0.830	0.858
WESAD (BVP-HRV) RFC	0.768±0.146	0.826±0.090
WESAD (BVP-HRV) SVM	0.763±0.103	0.792±0.096
WESAD (BVP-HRV) MLP	**0.780**±0.149	**0.829**±0.104
Test on ForDigitStress, BVP		
ForDigitStress baseline [19]	0.784	0.797
ForDigitStress RFC	0.810±0.177	0.829±0.143
ForDigitStress SVM	0.787±0.191	0.811±0.154
ForDigitStress MLP	**0.814**±0.179	**0.831**±0.152
Test on VerBIO, BVP		
VerBIO baseline [43]	-	-
VerBIO RFC	**0.949**±0.144	**0.962**±0.107
VerBIO SVM	0.879±0.178	0.904±0.128
VerBIO MLP	0.924±0.183	0.939±0.146

Table 4: Cross-dataset evaluation of SWELL-KW models on WESAD (ECG-HRV) dataset

Model	F1-score	Accuracy
Test on WESAD (ECG-HRV)		
SWELL-KW RFC	**0.557**±0.021	**0.676**±0.010
SWELL-KW SVM	0.357±0.006	0.482±0.013
SWELL-KW MLP	0.427±0.047	0.574±0.073

The SWELL-KW models performed poorly in the cross-dataset evaluation using WESAD ECG-HRV data. The RFC models achieved the best average performance with an f1-score of 0.557 and an accuracy of 0.676. These values are considerably lower than the worst-performing within-dataset WESAD (ECG-HRV) model, which achieved f1-score = 0.814 and accuracy = 0.862.

Table 5: Cross-dataset evaluation of WESAD (ECG-HRV) models on SWELL-KW dataset

Model	F1-score	Accuracy
Test on SWELL-KW		
WESAD (ECG-HRV) RFC	0.432±0.005	0.450±0.004
WESAD (ECG-HRV) SVM	0.421±0.006	0.439±0.003
WESAD (ECG-HRV) MLP	**0.468**±0.027	**0.479**±0.025

Like SWELL-KW models, the WESAD ECG-HRV models also performed poorly in cross-dataset evaluation. While MLP achieved the highest average performance in this cross-dataset evaluation, all models yielded f1-scores and accuracies below 0.50.

We employed a similar cross-dataset evaluation approach for models trained on BVP-derived HRV features. Each LOSO model was tested on unseen data from two different datasets. For instance, the WESAD (BVP-HRV) models were evaluated using data from the ForDigitStress and VerBIO datasets. The results of cross dataset evaluation for WESAD (BVP-HRV), ForDigitStress, and VerBIO models are presented in Tables 6, 7, and 8, respectively.

Table 6: Cross-dataset evaluation of WESAD (BVP-HRV) models on ForDigitStress and VerBIO datasets

Model	F1-score	Accuracy
Test on ForDigitStress		
WESAD (BVP-HRV) RFC	0.731±0.087	0.740±0.072
WESAD (BVP-HRV) SVM	**0.774**±0.003	**0.775**±0.003
WESAD (BVP-HRV) MLP	0.740±0.032	0.744±0.029
Test on VerBIO		
WESAD (BVP-HRV) RFC	**0.904**±0.008	**0.907**±0.009
WESAD (BVP-HRV) SVM	0.873±0.003	0.877±0.003
WESAD (BVP-HRV) MLP	0.888±0.004	0.893±0.003

The models trained on WESAD BVP-HRV features showed good cross-dataset performance on ForDigitStress and VerBIO datasets. SVM achieved the best average f1-score and accuracy on the ForDigitStress dataset, whereas RFC was better in the VerBIO dataset. The performance drop compared to the best within-dataset models on ForDigitStress and VerBIO was minimal, ranging from 0.04 to 0.06.

Table 7: Cross-dataset evaluation of ForDigitStress models on WESAD (BVP-HRV) and VerBIO datasets

Model	F1-score	Accuracy
Test on WESAD (BVP-HRV)		
ForDigitStress RFC	**0.789**±0.003	**0.820**±0.003
ForDigitStress SVM	0.779±0.002	0.812±0.003
ForDigitStress MLP	0.763±0.027	0.810±0.023
Test on VerBIO		
ForDigitStress RFC	**0.856**±0.004	**0.865**±0.005
ForDigitStress SVM	0.836±0.007	0.846±0.006
ForDigitStress MLP	0.784±0.049	0.805±0.038

The ForDigitStress models performed well the WESAD (BVP-HRV) and VerBIO datasets, with all models achieving f1-scores higher than 0.75 and more than 0.80 accuracy. The RFC showed the best performance on both external datasets (pvalue > 0.1).

The VerBIO models exhibited good generalizability to unseen data from both WESAD (BVP-HRV) and ForDigitStress datasets. A notable exception was the SVM models tested on ForDigitStress achieved f1-scores slightly below 0.70, representing a drop of 0.10 - 0.12 compared to the lowest performing within-dataset ForDigitStress model. Interestingly, the VerBIO MLP - which performed the best on ForDigitStress data - achieved a slightly better performance than the within-dataset ForDigitStress models.

Table 8: Cross-dataset evaluation of VerBIO models on WESAD (BVP-HRV) and ForDigitStress datasets

Model	F1-score	Accuracy
Test on WESAD (BVP-HRV)		
VerBIO RFC	0.744±0.001	0.759±0.001
VerBIO SVM	0.769±0.003	0.789±0.003
VerBIO MLP	**0.780**±0.035	**0.803**±0.036
Test on ForDigitStress		
VerBIO RFC	**0.823**±0.001	**0.823**±0.001
VerBIO SVM	0.684±0.010	0.688±0.009
VerBIO MLP	0.747±0.081	0.752±0.073

4.3 Combining Datasets

To explore the potential benefits of combining datasets, we trained additional models using the LOSO technique. The SWELL-KW and WESAD (ECG-HRV) data were combined to train ECG-derived HRV models. Similarly, a combined dataset consisting of WESAD (BVP-HRV), ForDigitStress, and VerBIO data was used to train BVP-derived HRV models. The LOSO results for combined ECG-based and BVP-based HRV models are presented in Tables 9 and 10, respectively.

Table 9: Results of LOSO evaluation of ECG-derived HRV models trained by combining data from SWELL-KW and WESAD (ECG-HRV) datasets

Model	F1-score	Accuracy
Test on SWELL-KW		
Combined ECG-HRV RFC	0.644±0.157	0.671±0.143
Combined ECG-HRV SVM	0.587±0.160	0.615±0.145
Combined ECG-HRV MLP	**0.660**±0.147	**0.679**±0.141
Test on WESAD (ECG-HRV)		
Combined ECG-HRV RFC	0.679±0.205	0.761±0.141
Combined ECG-HRV SVM	0.526±0.213	0.652±0.159
Combined ECG-HRV MLP	**0.718**±0.212	**0.792**±0.156
Combined ECG-HRV Results		
Combined ECG-HRV RFC	0.658±0.179	0.707±0.149
Combined ECG-HRV SVM	0.587±0.185	0.615±0.152
Combined ECG-HRV MLP	**0.683**±0.179	**0.725**±0.157

Among the models trained on combined SWELL-KW and WESAD (ECG-HRV), MLP outperformed others on both datasets (pvalue > 0.1). However, the performances on each dataset were lower than the within-dataset models. While the drop in performance for the SWELL-KW dataset was relatively small, there was around 0.12 drop in f1-score and 0.09 drop in accuracy for the WESAD dataset.

Combining WESAD (BVP-HRV), ForDigitStress, and VerBIO datasets resulted in models with good stress detection performance across all three datasets. The best average f1-scores and accuracies for each dataset were greater than 0.80. MLP trained on the combined dataset outperformed the best within-dataset WESAD (BVP-HRV) model by around 0.04. The best performance of combined models on the ForDigitStress dataset was similar to the within-dataset results. However, in the VerBIO dataset, the highest average

Table 10: Results of LOSO evaluation of BVP-derived HRV models trained by combining data from WESAD (BVP-HRV), ForDigitStress, and VerBIO datasets

Model	F1-score	Accuracy
Test on WESAD (BVP-HRV)		
Combined BVP-HRV RFC	0.785±0.106	0.816±0.085
Combined BVP-HRV SVM	0.768±0.138	0.813±0.096
Combined BVP-HRV MLP	**0.823**±0.131	**0.863**±0.087
Test on ForDigitStress		
Combined BVP-HRV RFC	0.809±0.172	0.828±0.139
Combined BVP-HRV SVM	0.776±0.183	0.800±0.151
Combined BVP-HRV MLP	**0.811**±0.185	**0.831**±0.150
Test on VerBIO		
Combined BVP-HRV RFC	**0.909**±0.177	**0.934**±0.121
Combined BVP-HRV SVM	0.845±0.212	0.886±0.149
Combined BVP-HRV MLP	0.880±0.203	0.913±0.146
Combined BVP-HRV Results		
Combined BVP-HRV RFC	**0.850**±0.175	0.**874**±0.135
Combined BVP-HRV SVM	0.806±0.194	0.840±0.149
Combined BVP-HRV MLP	0.844±0.190	0.873±0.145

performance was 0.03 - 0.04 lower than the best within-dataset performance.

5 Discussion

The MLP models achieved the best results in most within-dataset evaluations. Our observation aligns with the findings of Bobade and Vani [10], Prajod and André [32], and Albaladejo-González et al. [2]. They observed that a simple feed-forward network achieved better performance than other machine learning methods such as SVM and RFC. Additionally, we observed that RFC performed better in many cross-dataset evaluations. However, this trend is not very consistent across datasets.

Cross-dataset evaluations revealed significant limitations in generalizability for models trained on SWELL-KW and WESAD (ECG-HRV) datasets. Combining these datasets did not improve performance and, in the case of WESAD models, even led to a decline. This observation highlights the importance of data compatibility when considering such strategies. Our findings regarding the cross-dataset performance and combining datasets with respect to SWELL-KW and WESAD datasets are in line with the observations of previous works [2, 32].

As highlighted in Table 2, the SWELL-KW and WESAD (ECG-HRV) datasets differ in many factors including stressors, experienced stress intensity, and the measurement devices. To pinpoint the dataset characteristics that considerably influence the cross-dataset performance, we considered two additional datasets: ForDigitStress and VerBIO. The models trained on these three datasets (WESAD, VerBIO, ForDigitStress) exhibited good cross-dataset performance. This suggests that factors like the brand of measurement device, stress-elicitation technique, and even stress intensity (within a reasonable range) may not significantly impact generalizability when the primary stressor remains consistent (social evaluation in this case). Two studies from the reviewed literature - Mishra et al. [27] and Baird et al. [5] - observed good cross-dataset performances. However, both the studies compared stress datasets with many similarities, and especially utilized virtually identical tasks (mental arithmetic in Mishra et al. and TSST in Baird et al.). Together with their findings, we infer that the primary stressor plays a crucial role in the cross-dataset generalizability of stress models.

It is important to note that the investigated datasets were limited in terms of certain factors such as stress intensity and environmental aspects (e.g., temperature), and thus, these factors need further exploration. Moreover, our study focused on ECG- and BVP-derived HRV features. In future works, we will explore whether these findings extend to other stress-related modalities (e.g., EDA). Additionally, a more extensive evaluation involving a wider range of datasets encompassing other stressors (e.g., physical stress) would further validate these observations. Another aspect to investigate is the applicability of these models to natural stressful situations (e.g., social stress during musical performances).

Our research focused on developing stress detection models with broader applicability to various HMI systems. Based on our results, we suggest matching the primary stressor between the training dataset and the target deployment scenario for better cross-dataset generalizability. Moreover, combining datasets was beneficial only when the stressor matched. In other cases, techniques such as domain generalization and domain adaptation could be adopted to improve generalizability. For instance, recently such techniques have been used in the context of depression detection [42] and suicidal ideation [31]. However, these techniques are relatively unexplored in the context of stress detection, especially for datasets with different types of stressors. Nevertheless, assessing the compatibility of datasets is crucial before choosing between combining datasets and domain generalization/adaptation techniques.

6 Conclusion

Stress detection models with broader applicability are crucial due to the diverse nature of stress experiences across various scenarios. Identifying factors that influence cross-dataset generalizability is essential for achieving this goal. This study addressed this gap by conducting cross-dataset evaluations on four datasets (SWELL-KW, WESAD, ForDigitStress, VerBIO), which contain both shared and distinct characteristics. We trained HRV-based machine learning models (RFC, SVM, MLP) using ECG or BVP signals from these datasets. Our key finding is that the primary stressor is the most prominent factor influencing cross-dataset applicability. Models trained on datasets with similar stressors exhibited good generalizability, while those with different stressors showed lower performance. Conversely, factors like stress elicitation method and stress intensity had a minimal impact within the explored datasets. Furthermore, matching primary stressors proved crucial for enhancing stress detection performance when combining datasets.

Acknowledgments

This work was carried out within the framework of the AI Production Network Augsburg and was partially funded by Deutsche Forschungsgemeinschaft (DFG) under the LEAF project (Project Number 471909988).

References

[1] Ayten Ozge Akmandor and Niraj K Jha. 2017. Keep the stress away with SoDA: Stress detection and alleviation system. *IEEE Transactions on Multi-Scale Computing Systems* 3, 4 (2017), 269–282.

[2] Mariano Albaladejo-González, José A Ruipérez-Valiente, and Félix Gómez Mármol. 2023. Evaluating different configurations of machine learning models and their transfer learning capabilities for stress detection using heart rate. *Journal of Ambient Intelligence and Humanized Computing* 14, 8 (2023), 11011–11021.

[3] Ane Alberdi, Asier Aztiria, and Adrian Basarab. 2016. Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. *Journal of biomedical informatics* 59 (2016), 49–75.

[4] Karim Assi, Lakmal Meegahapola, William Droz, Peter Kun, Amalia De Götzen, Miriam Bidoglia, Sally Stares, George Gaskell, Altangerel Chagnaa, Amarsanaa Ganbold, et al. 2023. Complex daily activities, country-level diversity, and smartphone sensing: A study in denmark, italy, mongolia, paraguay, and uk. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–23.

[5] Alice Baird, Andreas Triantafyllopoulos, Sandra Zänkert, Sandra Ottl, Lukas Christ, Lukas Stappen, Julian Konzok, Sarah Sturmbauer, Eva-Maria Meßner, Brigitte M Kudielka, et al. 2021. An evaluation of speech-based recognition of emotional and physiological markers of stress. *Frontiers in Computer Science* 3 (2021), 750284.

[6] Luke Balcombe and Diego De Leo. 2022. Human-computer interaction in digital mental health. In *Informatics*, Vol. 9. MDPI, 14.

[7] Mouna Benchekroun, Dan Istrate, Vincent Zalc, and Dominique Lenne. 2022. A Multi-Modal Dataset (MMSD) for Acute Stress Bio-Markers. In *International Joint Conference on Biomedical Engineering Systems and Technologies*. Springer, 377–392.

[8] Mouna Benchekroun, Pedro Elkind Velmovitsky, Dan Istrate, Vincent Zalc, Plinio Pelegrini Morita, and Dominique Lenne. 2023. Cross dataset analysis for generalizability of HRV-based stress detection models. *Sensors* 23, 4 (2023), 1807.

[9] Javad Birjandtalab, Diana Cogan, Maziyar Baran Pouyan, and Mehrdad Nourani. 2016. A non-EEG biosignals dataset for assessment and visualization of neurological status. In *2016 IEEE International Workshop on Signal Processing Systems (SiPS)*. IEEE, 110–114.

[10] Pramod Bobade and M Vani. 2020. Stress detection with machine learning and deep learning using multimodal physiological data. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 51–57.

[11] Jason J Braithwaite, Derrick G Watson, Robert Jones, and Mickey Rowe. 2013. A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology* 49, 1 (2013), 1017–1034.

[12] Yekta Said Can, Bert Arnrich, and Cem Ersoy. 2019. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of biomedical informatics* 92 (2019), 103139.

[13] Mohamed Elgendi, Mirjam Jonkman, and Friso De Boer. 2010. Frequency Bands Effects on QRS Detection. *Biosignals* 2003 (2010), 2002.

[14] Mohamed Elgendi, Ian Norton, Matt Brearley, Derek Abbott, and Dale Schuurmans. 2013. Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions. *PloS one* 8, 10 (2013), e76585.

[15] Giorgos Giannakakis, Dimitris Grigoriadis, Katerina Giannakaki, Olympia Simantiraki, Alexandros Roniotis, and Manolis Tsiknakis. 2019. Review on psychological stress detection using biosignals. *IEEE transactions on affective computing* 13, 1 (2019), 440–460.

[16] Shalom Greene, Himanshu Thapliyal, and Allison Caban-Holt. 2016. A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health. *IEEE Consumer Electronics Magazine* 5, 4 (2016), 44–56.

[17] Rohit Gupta, Amit Bhongade, and Tapan Kumar Gandhi. 2023. Multimodal Wearable Sensors-based Stress and Affective States Prediction Model. In *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Vol. 1. IEEE, 30–35.

[18] Yeaminul Haque, Rahat Shahriar Zawad, Chowdhury Saleh Ahmed Rony, Hasan Al Banna, Tapotosh Ghosh, M Shamim Kaiser, and Mufti Mahmud. 2024. State-of-the-Art of Stress Prediction from Heart Rate Variability Using Artificial Intelligence. *Cognitive Computation* 16, 2 (2024), 455–481.

[19] Alexander Heimerl, Pooja Prajod, Silvan Mertes, Tobias Baur, Matthias Kraus, Ailin Liu, Helen Risack, Nicolas Rohleder, Elisabeth André, and Linda Becker. 2023. ForDigitStress: A multi-modal stress dataset employing a digital job interview scenario. *arXiv preprint arXiv:2303.07742* (2023).

[20] Saskia Koldijk, Mark A Neerincx, and Wessel Kraaij. 2016. Detecting work stress in offices by combining unobtrusive sensors. *IEEE Transactions on affective computing* 9, 2 (2016), 227–239.

[21] Saskia Koldijk, Maya Sappelli, Suzan Verberne, Mark A Neerincx, and Wessel Kraaij. 2014. The SWELL knowledge work dataset for stress and user modeling research. In *Proceedings of the 16th international conference on multimodal interaction*. 291–298.

[22] Alexandros Liapis, Evanthia Faliagka, Christos Katsanos, Christos Antonopoulos, and Nikolaos Voros. 2021. Detection of subtle stress episodes during UX evaluation: Assessing the performance of the WESAD bio-signals dataset. In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part III 18*. Springer, 238–247.

[23] Tiffany Luong, Nicolas Martin, Anais Raison, Ferran Argelaguet, Jean-Marc Diverrez, and Anatole Lécuyer. 2020. Towards real-time recognition of users mental workload using integrated physiological sensors into a VR HMD. In *2020 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE, 425–437.

[24] Dominique Makowski, Tam Pham, Zen J Lau, Jan C Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and SH Annabel Chen. 2021. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior research methods* (2021), 1–8.

[25] Miguel Martinho, Ana Fred, and Hugo Silva. 2018. Towards continuous user recognition by exploring physiological multimodality: An electrocardiogram (ECG) and blood volume pulse (BVP) approach. In *2018 International Symposium in Sensing and Instrumentation in IoT Era (ISSI)*. IEEE, 1–6.

[26] Lakmal Meegahapola, William Droz, Peter Kun, Amalia De Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, et al. 2023. Generalization and personalization of mobile sensing-based mood inference models: an analysis of college students in eight countries. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 6, 4 (2023), 1–32.

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

[28] Marta Mondellini, Pooja Prajod, Matteo Lavit Nicora, Mattia Chiappini, Ettore Micheletti, Fabio Alexander Storm, Rocco Vertechy, Elisabeth André, and Matteo Malosio. 2023. Behavioral patterns in robotic collaborative assembly: comparing neurotypical and Autism Spectrum Disorder participants. *Frontiers in Psychology* 14 (2023), 1245857.

[29] Kizito Nkurikiyeyezu, Anna Yokokubo, and Guillaume Lopez. 2019. The effect of person-specific biometrics in improving generic stress predictive models. *arXiv preprint arXiv:1910.01770* (2019).

[30] Tam Pham, Zen Juen Lau, SH Annabel Chen, and Dominique Makowski. 2021. Heart rate variability in psychology: A review of HRV indices and an analysis tutorial. *Sensors* 21, 12 (2021), 3998.

[31] Arvind Pillai, Subigya Kumar Nepal, Weichen Wang, Matthew Nemesure, Michael Heinz, George Price, Damien Lekkas, Amanda C Collins, Tess Griffin, Benjamin Buck, et al. 2024. Investigating Generalizability of Speech-based Suicidal Ideation Detection Using Mobile Phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–38.

[32] Pooja Prajod and Elisabeth André. 2022. On the Generalizability of ECG-based Stress Detection Models. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 549–554.

[33] Pooja Prajod, Tobias Huber, and Elisabeth André. 2022. Using explainable AI to identify differences between clinical and experimental pain detection models based on facial expressions. In *International Conference on Multimedia Modeling*. Springer, 311–322.

[34] Pooja Prajod, Matteo Lavit Nicora, Matteo Malosio, and Elisabeth André. 2023. Gaze-based attention recognition for human-robot collaboration. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*. 140–147.

[35] Pooja Prajod, Matteo Lavit Nicora, Marta Mondellini, Matteo Meregalli Falerni, Rocco Vertechy, Matteo Malosio, and Elisabeth André. 2024. Flow in human-robot collaboration—multimodal analysis and perceived challenge detection in industrial scenarios. *Frontiers in Robotics and AI* 11 (2024), 1393795.

[36] Rita Meziati Sabour, Yannick Benezeth, Pierre De Oliveira, Julien Chappe, and Fan Yang. 2021. Ubfc-phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing* 14, 1 (2021), 622–636.

[37] Pritam Sarkar and Ali Etemad. 2020. Self-supervised ECG representation learning for emotion recognition. *IEEE Transactions on Affective Computing* 13, 3 (2020), 1541–1554.

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

[39] Pedro Elkind Velmovitsky, Paulo Alencar, Scott T Leatherdale, Donald Cowan, and Plinio Pelegrini Morita. 2021. Towards real-time public health: a novel mobile health monitoring system. In *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 6049–6051.

[40] Gideon Vos, Kelly Trinh, Zoltan Sarnyai, and Mostafa Rahimi Azghadi. 2023. Ensemble machine learning model trained on a new synthesized dataset generalizes well for stress prediction using wearable devices. *Journal of Biomedical*

Informatics 148 (2023), 104556.

[41] Gideon Vos, Kelly Trinh, Zoltan Sarnyai, and Mostafa Rahimi Azghadi. 2023. Generalizable machine learning for stress monitoring from wearable devices: a systematic literature review. *International Journal of Medical Informatics* 173 (2023), 105026.

[42] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S Kuehn, Jeremy F Huckins, Margaret E Morris, et al. 2023. GLOBEM: cross-dataset generalization of longitudinal human behavior modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–34.

[43] Megha Yadav, Md Nazmus Sakib, Ehsanul Haque Nirjhar, Kexin Feng, Amir H Behzadan, and Theodora Chaspari. 2020. Exploring individual differences of public speaking anxiety in real-life and virtual presentations. *IEEE Transactions on Affective Computing* 13, 3 (2020), 1168–1182.

[44] Bin Yu, Mathias Funk, Jun Hu, Qi Wang, and Loe Feijs. 2018. Biofeedback for everyday stress management: A systematic review. *Frontiers in ICT* 5 (2018), 23.

[45] Md Rahat Shahriar Zawad, Chowdhury Saleh Ahmed Rony, Md Yeaminul Haque, Md Hasan Al Banna, Mufti Mahmud, and M Shamim Kaiser. 2023. A Hybrid Approach for Stress Prediction from Heart Rate Variability. In *Frontiers of ICT in Healthcare: Proceedings of EAIT 2022*. Springer, 111–121.