International Journal of Medical Informatics 173 (2023) 105026

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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

Review article

Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review

Gideon Vos [a], Kelly Trinh [a], Zoltan Sarnyai [b], Mostafa Rahimi Azghadi [a,*]

[a] College of Science and Engineering, James Cook University, James Cook Dr, Townsville, 4811, QLD, Australia
[b] College of Public Health, Medical, and Vet Sciences, James Cook University, James Cook Dr, Townsville, 4811, QLD, Australia

ARTICLE INFO

MSC:
68T01
92C99

Keywords:
Stress
Wearable sensor
Machine learning
Generalization

ABSTRACT

Introduction: Wearable sensors have shown promise as a non-intrusive method for collecting biomarkers that may correlate with levels of elevated stress. Stressors cause a variety of biological responses, and these physiological reactions can be measured using biomarkers including Heart Rate Variability (HRV), Electrodermal Activity (EDA) and Heart Rate (HR) that represent the stress response from the Hypothalamic-Pituitary-Adrenal (HPA) axis, the Autonomic Nervous System (ANS), and the immune system. While Cortisol response magnitude remains the gold standard indicator for stress assessment [1], recent advances in wearable technologies have resulted in the availability of a number of consumer devices capable of recording HRV, EDA and HR sensor biomarkers, amongst other signals. At the same time, researchers have been applying machine learning techniques to the recorded biomarkers in order to build models that may be able to predict elevated levels of stress.

Objective: The aim of this review is to provide an overview of machine learning techniques utilized in prior research with a specific focus on model generalization when using these public datasets as training data. We also shed light on the challenges and opportunities that machine learning-enabled stress monitoring and detection face.

Methods: This study reviewed published works contributing and/or using public datasets designed for detecting stress and their associated machine learning methods. The electronic databases of Google Scholar, Crossref, DOAJ and PubMed were searched for relevant articles and a total of 33 articles were identified and included in the final analysis. The reviewed works were synthesized into three categories of publicly available stress datasets, machine learning techniques applied using those, and future research directions. For the machine learning studies reviewed, we provide an analysis of their approach to results validation and model generalization. The quality assessment of the included studies was conducted in accordance with the IJMEDI checklist [2].

Results: A number of public datasets were identified that are labeled for stress detection. These datasets were most commonly produced from sensor biomarker data recorded using the Empatica E4 device, a well-studied, medical-grade wrist-worn wearable that provides sensor biomarkers most notable to correlate with elevated levels of stress. Most of the reviewed datasets contain less than twenty-four hours of data, and the varied experimental conditions and labeling methodologies potentially limit their ability to generalize for unseen data. In addition, we discuss that previous works show shortcomings in areas such as their labeling protocols, lack of statistical power, validity of stress biomarkers, and model generalization ability.

Conclusion: Health tracking and monitoring using wearable devices is growing in popularity, while the generalization of existing machine learning models still requires further study, and research in this area will continue to provide improvements as newer and more substantial datasets become available.

1. Introduction

Stress can be defined as the body's psychological and physiological response to physical, emotional or mental strain. Such change in the environment elicits the activation of a cascade of biological responses (stress response) in the brain and in the body [1,3]. The stress response serves an important evolutionary role of helping the adaptation of the organism to the dynamically changing external and internal

* Corresponding author.
E-mail address: mostafa.rahimiazghadi@jcu.edu.au (M. Rahimi Azghadi).

https://doi.org/10.1016/j.ijmedinf.2023.105026
Received 7 November 2022; Received in revised form 21 February 2023; Accepted 23 February 2023
Available online 28 February 2023
1386-5056/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

environment. This is achieved through mobilization of energy and its appropriate redistribution to organs that most immediately serve the adaptational response. At present, a universally recognized standard for stress evaluation remains outstanding [4], further compounded with the need for a comprehensive framework for investigating how organisms function in and adapt to constantly changing environments [5]. In the context of this paper and its reviewed studies, stress is considered as a binary condition for prediction. The data in a number of these studies [6,7] was labeled with binary stressed or non-stressed time periods, and models trained on these datasets resulted in classifiers that would predict an observation as either stressed or not-stressed, while the other datasets [8] utilized a daily stress inventory score [9] and one dataset [10] was labeled through observer scoring (0 to 1, low to high). A single study by Siirtola et al. [11] investigated and compared models trained as classifiers to models trained as a regression, where a threshold was established by analyzing the obtained continuous prediction values study subject-wise to obtain a balanced accuracy rate is as high as possible. In the studies reviewed, no single thresholding method could be determined that can generalize well across models.

Interestingly, a growing number of studies are examining the effects of training machine learning models on biomarker data collected in a study setting compared to daily life scenarios [12], with further studies examining the effect of context when both training and evaluating predictive power [13]. While the majority of studies in this review approached the training of machine learning models for stress detection as a single time-series dataset, more studies are evaluating the potential of person-specific models [14] compared to generic non-specific models, with person-specific models showing great promise as powerful predictors of stress.

Wearable devices for personal health monitoring and tracking have gained significant popularity and technical sophistication since the release of the first Fitbit [15] in 2009 and Empatica Embrace model in 2016 [16]. Recently, more advanced devices including Empatica's E4 [16] have been developed that are capable of measuring a wide variety of physiological signals. Peake et al. [17] performed a critical review of available wearable devices for providing bio-feedback, monitoring stress, and sleep with a critical review of their technical characteristics, reliability and validation. Continuous measurement of the physiological signals recorded using wearables enables researchers to extract useful information from these devices to potentially detect and monitor a variety of health-related events such as seizures [18–20], dehydration [21], cognitive load [22], physical activity [23], emotions [24] and specifically related to this review, stress [11,13,14,22,25–33].

A number of previous survey articles have studied the topics of stress detection using wearable devices [34] and machine learning [35]. In particular, in [34], Samson and Koh have surveyed various stress biomarkers and their measurement tools including wearables for salivary and electrochemical detection. However, they have not discussed how machine learning can be used to help with stress detection and measurements. In [35], Gedam and Paul have surveyed works that have performed stress detection using wearable sensors measuring Electrocardiogram (ECG), Electroencephalography (EEG), and Photoplethysmography (PPG) signals and surveyed machine learning techniques for that. However, in this paper, we systematically review studies that have mainly used biomarker data from medical-grade wearable devices available to the consumer, due to the growing popularity of personal health monitoring, different to those used in [35].

In addition, the previous reviews have not addressed a number of important points such as the statistical power [36] of the training data used or their labeling protocols, and how it may affect machine learning model performance. Neither have they considered machine learning model generalization, where models built on any of the available public stress datasets are capable of accurately measuring stress when applied on a new dataset, or applied on datasets recorded under different conditions including experimental set-up, session duration, and labeling methodology.

Towards addressing these questions, we first explore the current state of stress detection and measurement using medical-grade wearable devices that are available to the consumer. We further explore the available public datasets built using sensor data recorded from these devices, and investigate the approaches utilized, and detection accuracy scores attained for machine learning models trained on these datasets. Finally, we discuss the generalization ability and limitations of these machine learning models, in order to understand the current state of using wearable devices for accurately measuring stress response and future directions.

2. Methods

2.1. Research questions

The main aim of this work is to provide an overview of the current state of stress detection using machine learning techniques by using the IJMEDI checklist to assess the quality of the included literature, and specifically the generalization ability of models trained on public stress biomarker datasets and the potential reproducibility of their findings and results. Thus, our research questions can be formulated as follows:

- **RQ1:** Which machine learning algorithms and techniques are being utilized and trained on publicly available stress biomarker datasets?
- **RQ2:** What accuracy metrics are reported and how are these findings being validated? Are the findings reproducible and does the methods utilized show promise towards model generalization?

Answering these questions will aid in getting a better understanding of the most current and accurate machine learning models available for predicting stress using wearable devices, and assist towards building a model capable of generalization on new, unseen data.

2.2. Search strategy

We reviewed key published works (Fig. 1) between 2012 and 2022 on publicly available datasets related to stress, and more specifically, recorded using wearable devices; and measuring and predicting stress response using machine learning. The electronic databases of Google Scholar, Crossref, DOAJ and PubMed were searched for relevant articles using the keywords *stress*, *machine learning* and *wearable* in title or abstract, and a total of 973 papers were identified. Duplicates were identified, and 16 were found and removed, leaving the number of considered papers for the subsequent phases at 957. Abstracts were scanned and irrelevant papers were excluded, including papers where the full text was not available. A small number of papers, in which the focus was stress in animals or psychiatry, were excluded. Studies using devices that are generally considered as health-trackers, or lifestyle monitors were also excluded, as were studies performed solely using devices that would not generally be considered a wearable device, such as EEG or chest-worn monitors. We further limited this review to machine learning models trained on, and devices that are capable of, recording multiple biomarkers that are known to be robust indicators of elevated levels of stress, i.e. HRV, EDA, HR, Inter-beat Interval (IBI) [34]. Finally, papers where key machine learning techniques including feature-engineering and model validation techniques were not detailed, were also removed. As a result, a total of 33 papers were chosen for the systematic review process, grouped by the high-level topics of: *Datasets*, *Machine Learning for Stress Detection* and *Future Research and Open Problems*. Table 1 details the papers included in this review.

2.3. Assessment of the quality of the studies

Two reviewers (Vos and Azghadi) used the IJMEDI checklist [2] to evaluate the quality of the included studies independently. The IJMEDI checklist is a quality assessment tool for medical artificial intelligence

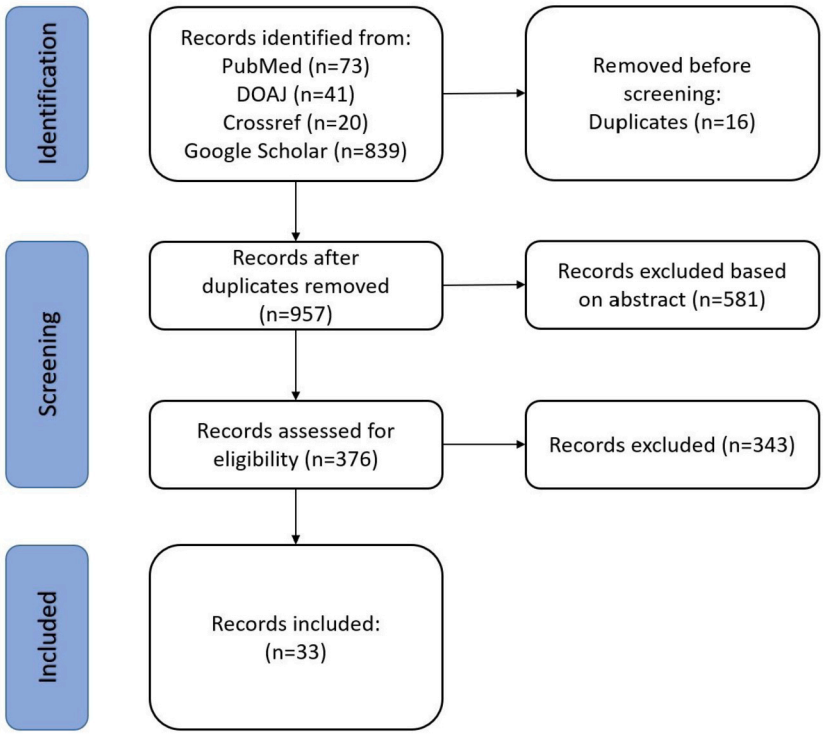

Fig. 1. Article screening process and the intermediate counts.

studies proposed by the IJMEDI, which aims to distinguish high-quality machine learning studies from simple medical data-mining studies. Six dimensions are included as 30 questions in the checklist: problem and data understanding, data preparation, modeling, validation, and deployment. Each question can be answered as OK (adequately addressed), mR (sufficient but improvable), and MR (inadequately addressed). In high-priority items, OK, mR and MR were assigned the scores of 0, 1, and 2, respectively, whereas in low priority items, the scores were halved. The maximum possible score was 50 points, with study quality was divided into low (0–19.5), medium (20–34.5), and high (35–50).

3. Results

3.1. Wearable devices for stress measurement

Advances in hardware such as component miniaturization have enabled more technological features to be embedded into ever shrinking devices at lower cost. However, adoption is clearly a challenge that demands the collaborative attention of healthcare providers, hardware and software engineers, data scientists, policy-makers, cognitive neuroscientists, device engineers and materials scientists, among other specializations [52]. From the initial Fitbit device launched in 2009, through to the Empatica E4 and the latest Oura Ring 3, significant improvements have been realized in both base features, as well as capabilities specifically related to the monitoring of, and promise to assist in improving, the user's overall health.

There are a wide variety of wearable devices in the market [17] used for health monitoring, including both medical-grade devices (Empatica Embrace Plus, Empatica E4, NOWATCH, Oura Ring) and consumer-oriented devices (Apple iWatch, Fitbit, Garmin, Samsung Gear). Consumer-oriented devices generally provide web-based platforms and smartphone applications for reporting various health statistics and levels of stress, with no ability to extract raw biomarker sensor recordings for scientific study, in contrast to medical-grade devices, such as the Empatica range, that provides full biomarker data download and additional support for researchers to properly utilize the raw signals directly for study.

However, in this review, our focus was limited to devices that are capable of stand-alone monitoring, without the need for an additional harness or pairing with a secondary device (worn on the wrist, finger or arm), as this would limit the usefulness for study outside of a stricter laboratory setting. Table 2 provides a non-exhaustive list of well-known wearable devices potentially capable of tracking and monitoring stress.

Siirtola [53] performed a study on smart watches reporting stress using a single biomarker (HR) and concluded that to be sufficient for detecting stress. Farrow et al. [54] concluded that EDA is a robust, reliable, non-subjective psycho-physiological biomarker of psychological stress within subjects, but not always between. Greco et al. [32] concluded that using only the EDA biomarker is sufficient for accurately predicting stress. The validity of sensor biomarkers is an open research question, discussed in detail in Section 4.1. Devices reporting stress based on only a single biomarker (typically HR or HRV) were therefor excluded.

The studies included in this review predominantly utilized datasets that are publicly available and therefor available to other researchers, and of these, the predominant wearable device utilized was the Empatica E4. Patient privacy when utilizing public health data for wearable research remains a concern, and Differential Privacy (DP) has emerged as a proficient technique to publish privacy sensitive data, including data from wearable devices. Saifuzzaman et al. [55] conducted a Systematic Literature Review to identify, select and critically appraise research in DP to understand the different techniques available in wearable data publishing, and proposed a number of solutions for protecting patient privacy. Of the public datasets reviewed and included in this study, all patient identifiable information were excluded from the datasets.

Additionally, the measurement of stress in people with mental disorders or intellectual disabilities is of growing interest. Simons et al. [56] presented a specific protocol for studying patterns of physiological stress in patients with challenging behavior. However, in this review we found the vast majority of current studies were performed using data captured from predominantly healthy subjects, screened for a number of health conditions prior to inclusion. Table 3 lists the studies included in this review where health screening was explicitly noted in the study,

Table 1
Studies included in this review.

Topic	Reference	Paper	Date
Data Sets	[7]	The swell knowledge work dataset for stress and user modeling research	2015
Data Sets	[6]	Introducing WESAD, a multimodal dataset for wearable stress and affect detection	2018
Data Sets	[10]	AffectiveROAD system and database to assess driver's attention	2018
Data Sets	[22]	Datasets for cognitive load inference using wearable sensors and psychological traits	2020
Data Sets	[8]	Multilevel monitoring of activity and sleep in healthy people	2020
Data Sets	[37]	K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations	2020
Data Sets	[38]	Toadstool: A dataset for training emotional intelligent machines playing super mario bros	2020
Data Sets	[39]	Comparison of machine learning techniques for psycho-physiological stress detection	2022
Machine Learning	[40]	Comparison of machine learning techniques for psycho-physiological stress detection	2016
Machine Learning	[13]	Monitoring stress with a wrist device using context	2017
Machine Learning	[29]	Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study	2019
Machine Learning	[25]	Objective Measurement of Physician Stress in the Emergency Department Using a Wearable Sensor	2020
Machine Learning	[30]	Stress Detection from Multimodal Wearable Sensor Data	2020
Machine Learning	[33]	Detection and Characterization of Physical Activity and Psychological Stress from Wristband Data	2020
Machine Learning	[41]	Evaluating the Reproducibility of Physiological Stress Detection Models	2020
Machine Learning	[23]	Deep neural networks for human activity recognition with wearable sensors: Leave-one- subject-out cross-validation for model selection	2020
Machine Learning	[42]	Predicting stress in teens from wearable device data using machine learning methods	2020
Machine Learning	[11]	Comparison of regression and classification models for user-independent and personal stress detection	2020
Machine Learning	[22]	Datasets for cognitive load inference using wearable sensors and psychological traits	2020
Machine Learning	[43]	Cognitive Training and Stress Detection in MCI Frail Older People Through Wearable Sensors and Machine Learning	2020
Machine Learning	[12]	How Laboratory Experiments Can Be Exploited for Monitoring Stress in the Wild: A Bridge Between Laboratory and Daily Life	2020
Machine Learning	[44]	Machine Learning Ranks ECG as an Optimal Wearable Biosignal for Assessing Driving Stress	2020
Machine Learning	[45]	Stress Detection by Machine Learning and Wearable Sensors	2021
Machine Learning	[46]	HRV Features as Viable Physiological Markers for Stress Detection Using Wearable Devices	2021
Machine Learning	[28]	A Sensitivity Analysis of Biophysiological Responses of Stress for Wearable Sensors in Connected Health	2022
Machine Learning	[27]	An Advanced Stress Detection Approach based on Processing Data from Wearable Wrist Devices	2021
Machine Learning	[35]	A review on mental stress detection using wearable sensors and machine learning techniques	2021
Machine Learning	[32]	Acute stress state classification based on electrodermal activity modeling	2021
Machine Learning	[47]	Advancing Stress Detection Methodology with Deep Learning Techniques Targeting UX Evaluation in AAL Scenarios	2021
Machine Learning	[48]	A Conditional GAN for Generating Time Series Data for Stress Detection in Wearable Physiological Sensor Data	2022
Machine Learning	[49]	Exploring Unsupervised Machine Learning Classification Methods for Physiological Stress Detection	2022
Future Research	[50]	The quantified self: Fundamental disruption in big data science and biological discovery	2013
Future Research	[51]	Deriving a cortisol-related stress indicator from wearable skin conductance measurements: Quantitative model experimental validation	2020

or where a dataset was utilized that was built using biomarker data from subjects screened for inclusion based on reported health status.

3.2. Wearable device datasets for stress measurement

A number of datasets are publicly available containing sensor data recorded using a variety of devices matching our inclusion criteria, as detailed in Table 4. The reviewed datasets contain the biomarkers predominantly utilized for stress detection, specifically EDA and HR signals. Apart from the Toadstool dataset, all recorded sessions exceed 60 minutes. The AffectiveROAD and Toadstool datasets contain biomarkers for a relatively small sample size of 10 subjects each, and small sample sizes of 25 subjects or less is a common feature of all public datasets reviewed. The largest public dataset included for review, Stress-Predict [39], contains biomarker data recorded using an Empatica E4 device for 35 test subjects.

Labeling of the included datasets were performed using one of two methods: (i) periodic, where specific time frames during the experiment were either labeled as stressed or non-stressed, while the test subject was placed under that perceived condition (a stressful test or action, or non-stressed, restful period), or (ii) scored as experiencing stress or no stress during a particular period, either by completing a self-scoring evaluation, or by an observer who perceived a level of stress by observing the emotional reaction of the subject during that period.

The American Psychological Association defines three types of stress - Acute, Episodic Acute and Chronic, further divided into Absolute Stressors (stressors that everyone exposed to them would interpret as being

Table 2
Wearable devices for health tracking and monitoring.

Device	Release Year	Type	Sensors	Battery Life
NOWATCH	2023	Wrist	HR, TEMP, SpO2, EDA	2 Weeks
Empatica Embrace Plus	2022	Wrist	EDA, ACC, TEMP, PR, PRV, ACC	1 Week
Fitbit Sense 2	2022	Wrist	HR, TEMP, SpO2, EDA	6 Days
Oura Ring 3	2021	Finger	HR, TEMP, SpO2, EDA	1 Week
Samsung Galaxy Watch 3	2020	Wrist	BVP, HR, ACC	48 Hours
Apple Watch 7	2019	Wrist	HR, ACC, SpO2	18 Hours
Fossil Gen 5	2019	Wrist	BVP, HR, ACC	24 Hours
Garmin Fenix 6X Pro	2019	Wrist	BVP, HR, ACC, SpO2	21 Days
Polar OH1	2019	Arm	BVP, ACC	12 Hours
Fitbit Charge 3	2018	Wrist	HR, ACC	1 Week
Garmin VivoActive 3	2018	Wrist	HR, ACC	1 Week
Study Watch	2017	Wrist	HR, TEMP, EDA	1 Week
Moodmetric	2017	Finger	EDA	1 Week
Empatica E4	2015	Wrist	HR, TEMP, SpO2, EDA, ACC, IBI	48 Hours
Sony SmartBand 2	2015	Wrist	BVP, HR, ACC	10 Hours
Samsung Gear Live	2014	Wrist	BVP, HR, ACC	24 Hours
Philips DTI-2	2014	Wrist	EDA, ACC, TEMP	30 Hours

ACC - accelerometer, BVP - Blood volume pulse, EDA - Electrodermal activity.
HR - Heart rate, IBI - Inter-beat Interval, PR - Pulse rate.
PRV - Pulse rate variability, SpO2 - Oxygen saturation, TEMP - Temperature.

Table 3
Reported health screening criteria of reviewed studies.

Paper	Dataset	Healthy	Pregnancy	Smoking	Caffeine	Alcohol	Mental Disorders	Cardiovascular Disease
[28]	WESAD	•	•	•			•	•
[14]	WESAD	•	•	•			•	•
[43]	Custom	•					•	
[48]	Custom	•						
[6]	WESAD	•	•	•			•	•
[47]	WESAD	•	•	•			•	•
[27]	WESAD	•	•	•			•	•
[32]	Custom	•		•	•	•		
[30]	WESAD	•	•	•			•	•
[26]	Custom	•					•	
[49]	SWELL	•						•
[45]	WESAD	•	•	•			•	•

Table 4
Summary of reviewed public wearable device stress-related datasets.

Dataset	Year	Subjects	Female	Male	Duration	Biomarkers	Devices	Labeling/Scoring
SWELL	2014	25	8	17	138 min	EDA, HRV, ECG	Facial expression, body postures, Mobi	Periodic: Neutral, Time Pressure, Interruptions
Neurological Status	2017	20			31 min	ACC,EDA, TEMP, HR, SPO2	Empatica E4	Periodic: Relax, Physical Stress, Emotional Stress, Relax, Emotional Stress, Relax
WESAD	2018	15	3	12	120 min	ACC, EDA, BVP, IBI, HR, TEMP, ECG, EMG, RESP	RespiBAN, Empatica E4	Periodic: Preparation, Baseline, Amusement, Stress, Meditation, Recovery
AffectiveROAD	2018	10	5	5	118 min	EDA, HR, TEMP	Empatica and Zephyr BioHarness 3.0 chest belt	Scored by observer
Toadstool	2020	10	5	5	50 min	ACC, EDA, BVP, IBI, HR, TEMP	Empatica E4	Periodic: Game play under time pressure
MMASH	2020	22			24 hrs	ACC, EDA, BVP, IBI, HR, TEMP, Cortisol	Empatica E4	Daily Stress Inventory value (DSI)
K-EmoCon	2020	32	12	20	120 min	ACC, EDA, BVP, IBI, HR, TEMP, EEG	Empatica E4, Polar H7 Bluetooth Heart Rate Sensor, NeuroSky MindWave Headset	Self-report and observer scoring
Stress-Predict	2022	35			60 mins	BVP, HR, RR	Empatica E4	Periodic

Table 5
Summary of reviewed studies and study stressors types applied.

Paper	Dataset	Time Stress	Anticipatory Stress	Situational Stress	Encounter Stress
Ehrhart 2022 [48]	Custom		•	•	
Iqbal 2022 [49]	SWELL	•	•		
Iqbal 2021 [28]	WESAD	•	•	•	•
Liapis 2021 [47]	WESAD	•	•	•	•
Alshamrani 2021 [27]	WESAD	•	•	•	•
Greco 2021 [32]	Custom	•	•	•	
Garg 2021 [45]	WESAD	•	•	•	•
Delmastro 2020 [43]	Custom		•	•	
Indikawati 2020 [30]	WESAD	•	•	•	•
Gjoreski 2020 [22]	Custom	•		•	
Sevil 2020 [33]	Custom	•	•	•	•
Kaczor 2020 [25]	Custom	•	•	•	•
Han 2020 [31]	Custom	•	•	•	•
Jin 2020 [42]	Custom	•	•	•	•
Can 2020 [12]	Custom	•	•	•	•
Nkurikiyeyezu 2019 [14]	WESAD	•	•	•	•
Can 2019 [29]	Custom	•	•	•	
Schmidt 2018 [6]	WESAD	•	•	•	•
Smets 2016 [26]	Custom	•		•	

stressful) and Relative Stressors (stressors that only some exposed to them would interpret as being stressful). Albrecht [57] further defined four common types of stress, namely Time Stress, Anticipatory Stress (concerns about future events), Situational Stress (situations that you have no control over) and Encounter Stress (worry about interacting with a certain person or group of people).

Table 5 provides a summary of the types of stressors applied during each study reviewed in this paper, as defined by Albrecht [57], with a number of studies including all four types within their study setting and protocol. All studies involved cognitive or work-related tasks under pressure, and as noted in Table 3, study subjects were screened for known health conditions in virtually all studies. Of the studies reviewed, three collected stress biomarker data during normal life conditions.

Jin et al. [42] provided Empatica E4 devices to study subjects after device use training, allowing subjects to utilize the device event marker to indicate periods during the day when they felt moderate to high levels of stress. Kaczor et al. [25] performed a similar study in an healthcare emergency department, while Can et al. [12] investigated the predictive performance of models trained under laboratory conditions when predicting on data collected in normal life conditions, and found that models trained on data recorded during laboratory sessions outperformed models trained on data collected during normal daily life conditions, when predicting for daily life conditions. This particular study [12] is of importance to researchers interested in building models from study data, for use on patient data collected during normal life conditions.

3.3. Machine learning algorithms and techniques for stress measurement using wearable data

Reviewing the literature, we found several machine learning techniques applied to detect elevated levels of stress using wearable devices. Table 6 lists the papers reviewed and the machine learning algorithms utilized. In the following subsections, we provide a discussion on the different steps of the machine learning pipelines utilized, and analyze how previous works have performed those steps, noting their strengths and limitations.

3.3.1. Pre-processing

Electronic sensors used in wearable devices for recording biomarkers differ widely, and subsequently operate and record on different sampling frequencies. For the Empatica E4, for instance, the EDA signal is sampled at 4 Hz, while the HR signal is sampled at 1 Hz. Recorded session data for both sensors will therefor differ in length, and researchers will have to pre-process the sensor data by down-sampling the EDA signal to 1 Hz to ensure a like for like timestamp match with the HR signal, and subsequently any stress metric label for the exact time period. In the studies reviewed, [14,22,30,42,48] specifically noted that down-sampling was applied on data used within their experiments.

Due to varying experimental protocols and the ease of collection of non-stress samples, data is likely to be unbalanced with more non-stress samples versus stressed samples present in any given dataset. Therefore, another usual pre-processing step performed on wearable stress data is class balancing that can be done in different ways. For instance, Nkurikiyeyezu et al. [14] balanced the recorded sensor data by randomly discarding some samples from the majority (non-stressed) class, and further applied logarithmic, square root, and Yeo-Johnson transformations to ensure a Gaussian distribution, as required by their use of a linear regression model. Can et al. [29] also performed class-balancing through random down-sampling of the majority class (non-stressed observations) to match the minority class (stressed observations).

As noted in Table 7, neither up-sampling nor down-sampling techniques showed a substantial difference or improvement in predictive power, and this may be due to the lack of a proven strategy employed when selecting which observations to discard, potentially causing information loss [58] of important biomarker data during the sampling process. Class balancing techniques all have varied benefits and risks, as noted in Table 7, and to this extent, a number of methods have been proposed to improve class-balancing re-sampling techniques. Deng et al. [59] proposed a unified approach for multivariate time series classification when data is imbalanced, while Lee et al. [60] used a semi-supervised technique known as Active Learning to mitigate the effect of imbalanced class labels. Jiang et al. [61] proposed a new oversampling method based on the classification contribution degree to deal with a number of shortcomings when using SMOTE (Synthetic Minority Oversampling Technique) [62], such as oversampling from noisy points. A notable drawback of reliance on class balancing when dealing with highly imbalanced datasets such as the stress biomarker datasets included in this study, where the stressed period is generally the minority class, is reproducibility and generalizability on new, unseen data that may contain significant outliers and a different class distribution, depending on the study setting and protocol used during biomarker recording. Further research is required to identify robust techniques for dealing with these class imbalances in physiological biomarker datasets.

Differences in data range, units and scale can be problematic for some machine learning algorithms and standardization is usually applied to scale the data to have a mean of 0 and a standard deviation of 1. Similarly, the goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. In the context of stress detection, normaliza-

Table 6
Summary of the machine learning models reviewed.

Paper	Year	Model	Dataset	Accuracy	Subjects	Features	Cross Validation	Window
[26]	2016	Bayesian networks	Custom	84.60%	20	22	LOSO	30 s, 29 s overlap
[13]	2017	SVM	Custom	71.00%	5	6	LOSO	6 min
[6]	2018	Random Forest, LDA, AdaBoost	WESAD	93.00%	15	82	LOSO	0.25 s, 5 s, 60 s
[29]	2019	Neural Network, Random Forest	Custom	97.92%	21	17	10-Fold	2 min, 20 min
[14]	2019	Random Forest, ExtraTrees	WESAD, SWELL	93.90%	15, 25	94	10-Fold	5 min, 10 min
[11]	2020	Bagged tree based ensemble	AffectiveROAD	82.30%	9	119	LOSO	60 s, 0.5 s overlap
[22]	2020	XGBoost	Snake, CogLoad	82.00%	23,23	44	LOSO	
[33]	2020	LDA	Custom	98.30%	24	2216	10-Fold	
[25]	2020	Naive Bays	Custom	64.50%	8	30	10-Fold	20 min
[31]	2020	K-nearest-neighbor	Custom	94.55%	17	25	10-Fold	60 s
[42]	2020	Random Forest	Custom	89.40%	8	756	10-Fold	
[30]	2020	RF	WESAD	92.00%	15	4	60/40 Split	0.25 s
[44]	2020	Ensemble	SRAD	75.02%	17	6		
[12]	2020	Random Forest	Custom	74.61%	14	2	10-Fold	
[43]	2020	Adaboost	Custom	85.30%	9	10	10-Fold	
[28]	2021	Logistic Regression	WESAD	85.71%	14	5	14-Fold	60 s
[32]	2021	SVM, RF	Custom	94.62%	65	14	LOSO	
[27]	2021	Neural Network	WESAD	85.00%	15	14	LOSO	
[45]	2021	RF	WESAD	83.34%	15	5	LOSO	10 s
[46]	2021	Gradient Boosting	SRAD	79.00%	17	7		
[47]	2021	Neural Network	WESAD	97.40%	15	36	UX Dataset	
[49]	2022	Random Forest	SRAD, SWELL	65.6%, 75%	17, 25	2	70/30 Split	
[48]	2022	Neural Network	Custom	72.62%	35		Test on unseen	16 s

Table 7
Class balancing methods employed in reviewed studies.

Paper	Class Balancing	Accuracy	Benefit	Risk
[33]	ADASYN [63]	98.30%	Reduces bias [63]	Generates minority outliers [64]
[29]	Upsampling of Minority Class	97.92%	Simple implementation [64]	Promotes overfitting [65]
[14]	Downsampling of Majority Class	93.90%	Simple implementation [64]	Information loss [58]
[42]	Random Sampling [58]	89.40%	Reduces bias [63]	Explainability
[43]	SMOTE [62]	85.30%	Surpasses random sampling methods [66]	Introduces noise, Overgeneralization [61]
[12]	Downsampling of Majority Class	74.61%	Simple implementation [64]	Information loss [58]
[48]	Data Augmentation	72.62%	Can reduce bias [67]	Explainability

tion and standardization were utilized by [14,22,28,33,46], with [22] experimenting on both raw and standardized data, and finding that standardization offered improved predictive performance across all 10 machine learning algorithms tested. Another usual pre-processing step on biomedical signals such as stress-related biomarkers collected by wearable devices is filtering. This is done to reduce outliers and any potential noise. For instance, [6] applied a 5 Hz low-pass filter on the raw EDA signal, [28] applied a high-pass filter on the raw EDA signal, while [14,27,31] applied a 4 Hz fourth-order Butterworth low-pass filter, followed by a moving average filter, to reduce outliers and remove noise from EDA sensor signals.

3.3.2. Feature-engineering

A common technique for extracting useful features representing physiological time series data, is to summarize the changing features of the existing data using summary statistics. Guo et al. [68] performed a study to evaluate summary statistics as features for clinical prediction tasks, and found that commonly used combinations of summary statistics such as [min, max, mean] and [min, max, mean, standard deviation (std)] achieved good prediction results in most cases. However, they reported that skew and kurtosis, which reflect the shape of a distribution, performed poorly when used individually as features for prediction, but appeared frequently in the optimal combinations, indicating that they can play a role as supplemental information.

The techniques noted by Guo et al. [68] were frequently applied in the stress detection studies reviewed. Fourteen of the reviewed approaches [6,11–14,26,28,29,31–33,44–46] utilized summary statistics of biomarkers using a sliding-window approach, ranging from 0.25 seconds in one experiment up to 20 minutes in others, with varying degrees of success. In [69], the author noted summary windows of 30 and 60 seconds are most often utilized, based on the hypothesis that this factor correlates with physiological response. Can et al. [29] decomposed the phasic and tonic components of the EDA signal using a convex optimization approach, as the tonic component includes more long-term slow changes, whereas phasic components include faster (event-related) changes. Both [29] and [22] found that sliding windows ranging between 10 and 17.5 minutes produced better detection accuracy, with [29] further noting that different machine learning algorithms relied on different window sizes, an important factor to consider for future research.

Jin et al. [42] used the *tsfresh* Python library to automatically generate 4536 features off their existing data and applied a Random Forest model as machine learning approach. To evaluate the performance of such a large number of features, the results were grouped around the key biomarkers (i.e. HR, EDA, TEMP), from which the features were engineered. Gjoreski et al. [13] used greedy step-wise selection to identify the top features considered most useful for their specific machine learning model, and further noted that when sensor-specific features are used, PPG-based features achieved higher predictive accuracy results, followed by the IBI and HR-based features. Iqbal et al. [28] found features based on HR and respiratory rate to be the most important, while Dalmeida et al. [46] focused their research specifically on HRV as a viable biomarker, and found HRV features to constitute good markers for stress detection.

3.3.3. Algorithm selection

Of the 23 machine learning based stress detection studies reviewed, we noted the use of 16 different algorithms, including combinations of Logistic Regression (LR), Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), Bayesian Networks (BN), Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), k-Nearest Neighbor (kNN), Multi-layer Perceptron (MLP), Multi-task

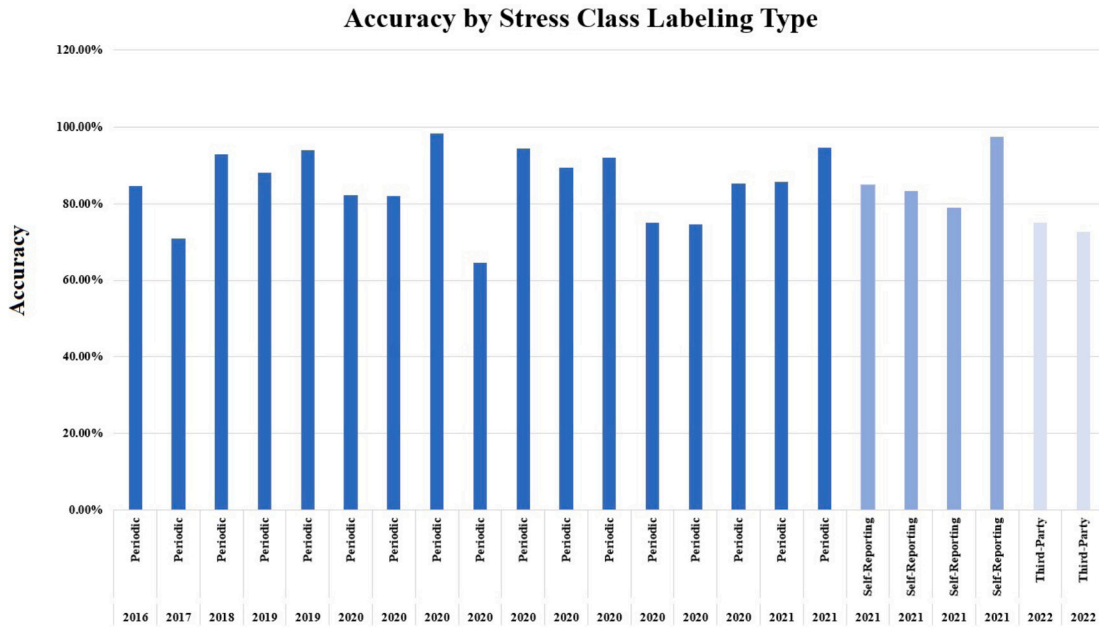

Fig. 2. Accuracy based on labeling method included in study.

learning (MTL), Adaboost, Naive Bayes (NB), Bagging, Gradient Boosting (GB) and Neural Networks (NN). Of these, SVM, RF and kNN were the most commonly used for stress detection, with tree-based models such as RF and GB generally delivering better predictive performance on supervised binary classification objectives.

A standard approach consists of selecting a small number of algorithms that may be suitable for the problem, train each and select the best performing model based on their final predictive accuracy. [14] experimented on a single method (Random Forest) while [53] used 13 different algorithms to test the predictive accuracy of classification based models versus regression type models for predicting elevated levels of stress, of which Bagged Trees performed the best. Similarly, [6,13,25,26,29,31] utilized 5 to 7 different algorithms and compared the stress prediction accuracy of each, with the highest performing models listed in Table 6. Iqbal et al. [49] compared the performance of 7 supervised methods to 7 unsupervised methods and concluded that a careful selection of classification models is required when aiming to develop an accurate stress detection system, with unsupervised machine learning classifiers showing good performance in terms of classification accuracy.

Additionally, the predictions from a set of algorithms can be combined based on averaging, weighted-averaging or voting, to produce a final prediction (commonly known as model ensembling). This technique was specifically noted in experiments done by Gjoreski et al. [13], Kaczor et al. [25] and Elgendi et al. [44].

3.3.4. Hyperparameter optimization

Hyperparameters can be defined as the different parameter values used to control the learning process of a machine learning algorithm, and can have a significant effect on their performance. Hyperparameter optimization is the process of finding the right combination of algorithm parameter values to achieve maximum performance on the given dataset. Examples of hyperparameters are the number of estimators (trees) and maximum tree depth in the Random Forest algorithm. Due to the large number of parameters that require tuning in different algorithms, automated methods [70] have been developed to scan the full parameter search space in a reasonable amount of time to determine the optimal combination.

Of the stress-related studies reviewed, we noted [26] restricted the hyperparameter of the estimators count used in their Random Forest model to 20, while [42] performed a grid search with estimators set at 500. In [14], the authors used 1,000 estimators while limiting the tree depth to 2, in order to limit the possibility of over-fitting. For the decision tree classification algorithms used by [6], information gain was used to measure the quality of splitting decision nodes, and the minimum number of samples required to split a node was set to 20. The number of base estimators was set to 100 for both of their utilized algorithms (Random Forest and AdaBoost). In another study, Han et al. [31] did not specifically optimize hyperparameters, but built several kNN models with different parameter values for k (1, 3, 5, 7, 9) and selected the best performing model from those. Sevil et al. [33] utilized Bayesian optimization techniques for feature-selection.

Unlike the aforementioned works, Gjoreski et al. [22] tuned their model parameters by randomly sampling from distributions predefined by an expert. The models were then trained with the specific parameters and evaluated using cross-validation on the training data. The best performing model from the cross-validation was used to classify the test data. A systematic, well-defined hyperparameter optimization approach is crucial to improve the reproducibility of scientific studies and ensures that machine learning algorithms are tailored to the problem at hand. As noted by Can et al. [29], the performance of machine learning models may be dependent on an optimal selection of window size when generating summary statistics to engineer features, and this needs consideration when selecting hyperparameters for optimal predictive performance.

3.3.5. Model training and validation

An important requirement when developing supervised machine learning algorithms is to have valid labeled data. In the case of stress measurement, we found three main methods employed for labeling elevated levels of stress. These include (i) specific stress/no-stress periods marked during an experimental recording session [6,7,25,27,29–33,38,39,43–49]; (ii) self-reporting via questionnaires [8,12,13,22,26,28,37]; and (iii) labeling by a third-party observer, who observes subjects' response to a situation and numerically scores/grades the level of stress observed [10,11,37,42]. Fig. 2 details the studies reviewed for each year, with reported accuracy rates by labeling method. Periodic labeling was the most commonly used labeling technique and provided consistently higher accuracy rates as reported by each study, compared to self-scoring and scoring by a third-party.

As highlighted in Table 6, the best performing models from each experiment achieved at least 64.5% test accuracy, with [6,14,30–33,47]

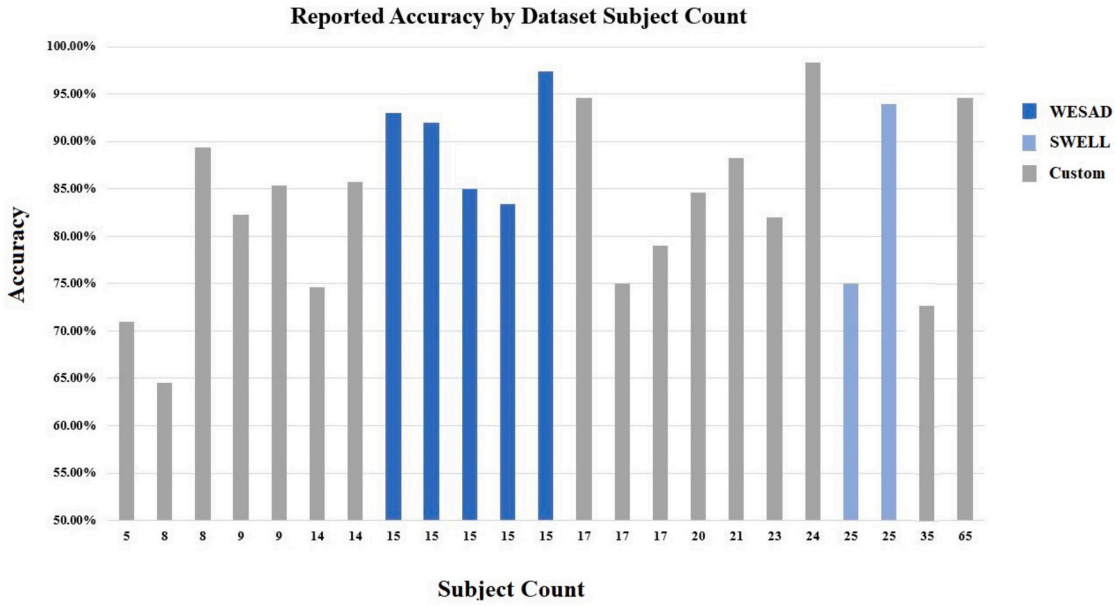

Fig. 3. Accuracy based on number of subjects included in study. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

reporting binary classification test accuracy rates of over 90%, using datasets labeled with specific, marked stress/no-stress periods. It should be noted that as stress is a physiological response, predictive accuracy in these experiments measures a predictive correlation between the included features (biomarkers) against a labeled metric at the same point in time (stressed versus non-stressed). Siirtola et al. [53] attempted to model how high this relationship is (using a regression algorithm instead of classification), while Umematsu et al. [71] focused on the problem of forecasting future episodes of stress, rather than measuring levels of stress on previously recorded data.

Cross-validation is a re-sampling procedure used to evaluate machine learning models on a limited data sample. The purpose of cross–validation is to test the ability of a machine learning model to predict with high accuracy on new, unseen data. It is also used to flag problems like over-fitting or selection bias, and gives insights on how well the model will generalize to an independent dataset. Among the studies reviewed, [6,11,13,22,26,27,32,45] utilized Leave One Subject Out (LOSO) cross-validation, while [12,14,25,28,29,31,33,42,43] utilized K-fold cross-validation with $K=10$. In addition, [22] utilized both LOSO and K-fold cross-validation, with $K=5$. All studies reviewed approached stress prediction as a binary classification problem apart from [53], where the problem type was defined as stress level measurement, rather than a binary stressed versus non-stressed problem. No definitive improvement in reported accuracy rates were noted when using LOSO cross-validation compared to K-fold cross-validation.

3.3.6. Performance analysis

A wide variety of metrics are available for measuring machine learning model performance, depending on the problem being solved, for example classification or regression type problems. The experiments reviewed utilized and reported a number of different evaluation metrics including F-score [6,22,29,45,48,49], classification accuracy [12,13,25–33,43–46,48,49], Kappa [47], Area Under the Curve (AUC) [42,46] and Mean Absolute Error (MAE) [14]. For comparisons among the reviewed studies, here we only investigate their achieved classification accuracy, if reported. Classification accuracy simply measures how often the classifier correctly predicts, i.e. what is the ratio of the number of correct predictions to the total number of predictions.

To determine classification accuracy, Kaczor et al. [25] divided their dataset into 3 classes: a pre-stress event vs post-stress event, baseline vs pre-stress event, and baseline vs post-stress event, reporting a classification accuracy rate of 90.4%. Can et al. [29] reported a binary classification accuracy rate of 97.92% using a custom (non-public) dataset, and Gjoreski et al. [22] reported a binary classification accuracy rate of 68.2% when applied on the *CogLoad* dataset and 82.3% when applied on the *Snake* dataset. Han et al. reported a binary accuracy rate of 94.55%, while Nkurikiyeyezu et al. [14] and Schmidt et al. [6] reported binary classification accuracy rates of 93.9% and 93% respectively. Jin et al. [42] reported an AUC rate of 89.4% rather than an accuracy rate. AUC, unlike classification accuracy, is sensitive to class imbalance when there is a minority class. This implies that classification accuracy rates can be high even if the predictions for a minority class are mostly wrong. This could lead to samples marked as non-stressed being classified mostly correctly and stressed samples (the minority class) predicted inaccurately, while still reporting an overall high accuracy rate.

Fig. 3 details the reported accuracy metrics achieved for the experiments reviewed in this paper, based on the size of the dataset used in terms of individual test subjects. The highest reported accuracy rate of 98.30% was achieved by Sevil et al. [33] when using Linear Discriminant Analysis (LDA) on a non-public dataset consisting of 24 test subjects, and validated using K-Fold cross-validation. Liapis et al. [47] trained a Neural Network on the public WESAD dataset (15 subjects), and evaluated on a user-annotated dataset consisting of skin conductivity (SC) segments for 30 study participants, reporting an accuracy rate of 97.40%.

3.3.7. Study quality

The *Supplementary File* details the results of the IJMEDI quality assessment. Table 8 summarizes the scores of each dimension and the total score in each study. The average score of the included studies was 25.7 (range: 14–35). Most of the studies were of a medium quality, while one [72] were of a high quality. The majority of the studies had an obvious bias in the quality of problem understanding, data understanding and modeling dimensions. Fig. 4 shows the proportion of the different answers in the high and low priority items.

4. Discussion

In order to build a robust machine learning model capable of accurately detecting stress, we consider four important requirements. These include (i) Sensor biomarker data needs to be valid and sufficiently

Table 8
Quality assessment scores of the 23 ML-based studies according to the IJMEDI checklist.

	Problem Understanding (10)	Data Understanding (6)	Data Preparation (8)	Modeling (6)	Validation (12)	Deployment (8)	Total (5)	Generalization
Iqbal 2021 [28]	7	6	8	6	7	2	36	
Nkurikiyeyezu 2019 [14]	6	3	8	6	8	2	33	
Delmastro 2020 [43]	9	5	2	6	7	4	33	
Gjoreski 2020 [22]	7	6	2	6	8	2	31	
Ehrhart 2022 [48]	7	5	2	6	9	2	31	•
Schmidt 2018 [6]	7	5	1	6	8	3	30	
Dalmeida 2021 [46]	6	4	4	6	8	2	30	
Liapis 2021 [47]	7	4	2	6	9	2	30	•
Can 2019 [29]	7	4	2	6	7	3	29	
Sevil 2020 [33]	5	4	4	6	7	2	28	
Alshamrani 2021 [27]	7	5	4	5	5.5	1	27.5	
Kaczor 2020 [25]	9	3	2	5	6	1	26	
Jin 2020 [42]	6	3	2	6	7	2	26	
Greco 2021 [32]	6	4	0	6	8	2	26	•
Indikawati 2020 [30]	7	5	2	6	5.5	0	25.5	
Can 2020 [12]	7	4	2	6	5.5	1	25.5	
Elgendi 2020 [44]	7	2	2	6	6	1	24	
Gjoreski 2017 [13]	5	3	2	5	5.5	3	23.5	
Smets 2016 [26]	6.5	5	0	5	5.5	1	23	
Han 2020 [31]	6.5	4	0	5	5.5	0.5	21.5	
Siirtola 2020 [11]	5	2	1	5	6	1	20	
Iqbal 2022 [49]	6	2	0	6	5	0	19	
Garg 2021 [45]	1	2	0	5	6	0	14	

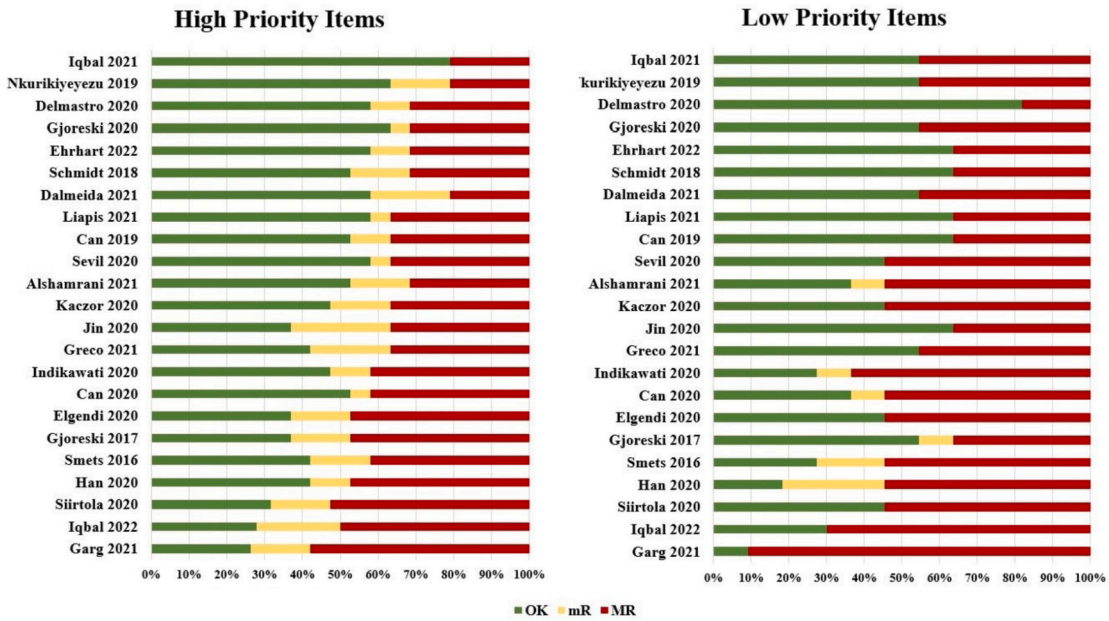

Fig. 4. Proportion of the different answers in the high- and low-priority items. OK = adequately addressed; mR = sufficient but improvable; MR = inadequately addressed.

varied to capture a wide spectrum of potential physiological stress response; (ii) For supervised machine learning, this data needs to be accurately labeled where observations are marked as stressed or non-stressed or a stress score range is given, to allow the model to learn from the data; (iii) Where a specific hypothesis is being tested, a sufficient level of statistical power is required [36,39,73–75], thereby ensuring results and findings can be considered statistically significant, and (iv) Model generalization occurs in order to apply the model on new, unseen data, with high accuracy. The discussion of this review is therefor focused on those four key requirements.

Having scored the machine learning studies included in this review using the IJMEDI checklist, we found only one study [72] of high quality, with the remaining studies being of medium quality, and a single study being of low quality [45]. Most studies scored well in problem, data understanding and modeling domains. Data preparation and deployment scores were notably low, as were scores for high priority items in the validation domain. Interestingly, there were no notable improvement in study quality over time.

Five studies [14,22,43,48,72] scored over 30, being of medium to higher quality. Focusing on the modeling, validation and deployment domain scores of these eight studies, we note an improvement in quality over time for only the validation domain, indicating a lack of progress in the modeling domain and little focus on the deployment of models in real-life scenarios, including factors pertaining to sustainability, model bias and ethics. See Fig. 5.

4.1. Validity of sensor biomarkers

The datasets included in this study contain a variety of sensor biomarker data potentially useful in detecting elevated levels of stress

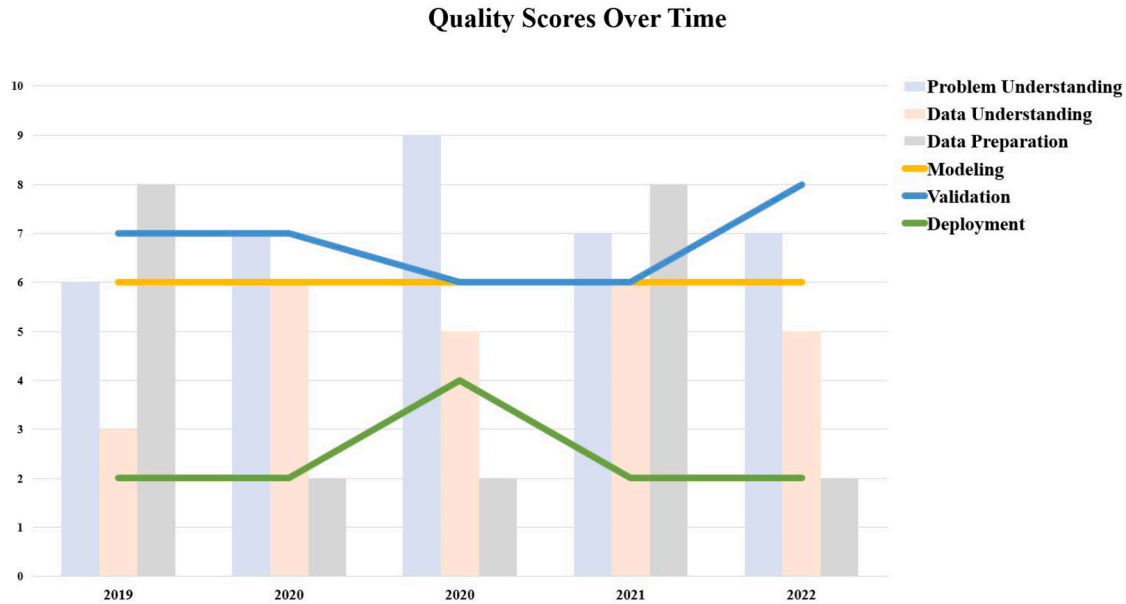

Fig. 5. Trend of higher-quality studies over time.

via HR, HRV and EDA signals, measured across a time interval. In addition, sweat sensing is at the forefront of wearable stress detection currently in development [34] and devices sensing sweat may hold great promise to quantify several biomarkers, namely Cortisol, to monitor the levels of stress that an individual is experiencing.

However, [76] noted that at present, there is a lack of consensus on a standardized protocol or framework with which to test the validity of physiological signals measured by these devices and their derived parameters. It is also argued that sudden, short-lived stressors, such as being startled by the ringing of the phone, or possible habituation effects as a result of exposure to repeated information cannot be validly detected. Lier et al. [76] reported that physiological changes during a workday can be tracked by the Empatica E4 wearable against more major, sustained stressors.

In [77], the authors found the Empatica E4 to be suitable for psychotherapy research focused on Inter-Beat Interval (IBI) and specific HRV measures, but failed to produce reliable EDA data and produced missing IBI data, especially when a subject is being more dynamic. This is confirmed by Ryan et al. [78] and Sevil et al. [33] that found the Empatica E4 can be severely compromised by motion artifact. This can result in a high percentage of missing data across all conditions except seated and supine baselines, and questions the E4's efficacy as an HRV measurement tool in most in-vivo conditions. This is further confirmed by Georgiou et al. [79] that found wearable devices can only be used as a surrogate for HRV at resting or mild exercise conditions, as their accuracy fades out with increasing exercise load, and Schuurmans et al. [80] who noted the potential of the Empatica E4 as a practical and valid tool for research on HR and HRV under non-movement conditions. Seipaejaervi et al. [81] found that an HRV-based stress index mirrors responses of cortisol, and an HRV-based stress index may be used to quantify physiological responses to psychosocial stress across various health and age groups. In contrast, Greco et al. [32] found EDA to be a good marker of stress when features are engineered based on its phasic and tonic components.

In [11], the authors predominantly focused on comparing regression vs. classification models using the AffectiveROAD dataset [10]. This dataset contains sensor recordings for both left and right hands of the test subjects. To ensure consistent, comparable results, [11] utilized only data recorded from the right hand of each test subject, leaving the important question of sensor placement unanswered, and needing further study to confirm whether sensor placement on the dominant versus non-dominant hand of a test subject could potentially affect biomarker accuracy, and more importantly for this review, correlation with increased levels of stress. Empatica note on their website [16] that newer studies have shown substantial differences in the EDA signal between the dominant and non-dominant hand.

4.2. Labeling protocol

In terms of labeling protocol and methodology, [11] questioned the accuracy of self-reporting of perceived levels of stress experienced, which was previously questioned by [82], who noted that study subjects are less likely to report on states less socially desired. Accurate labeling of stressed/non-stressed periods in the sensor data is crucial to building a reliable and robust machine learning model. To achieve this, in the datasets reviewed in Table 4, two major labeling methods were used. The SWELL, Toadstool and WESAD datasets were recorded with specific intervals to denote stressed/non-stressed periods for labeling. In the AffectiveROAD, MMASH and K-EmoCon datasets, on the other hand, labeling was performed using self or observed stress indicator scoring. An interesting observation is that where these datasets were utilized in reviewed machine learning models, the models trained on periodically-labeled data achieved significantly higher levels of detection accuracy compared to the models trained using self or observed stress scoring. This is likely due to false negative reporting in the questionnaires, as noted by [82].

Stress is not a binary condition, and none of the studies reviewed noted specific methods for establishing thresholds within biomarkers to utilize as indicators of periods of high stress (low, moderate, high). Any potential thresholds established by the machine learning algorithms, (specifically tree-based methods) during training were not examined in detail to determine any potential time-varying dynamics between the biomarkers. Ghiasi et al. [83] proposed combining HRV and EDA correlates as a single index, rather than treating each as separate indicators of ANS changes. They reported good results when validating this metric on two experimental protocols.

4.3. Lack of statistical power

Of the papers included in this review, six [14,26,28,43,44,46] specifically included a hypothesis statement in their experiments. However, no power analysis were noted in any of the machine learning papers reviewed, regardless of hypothesis statement. It is common to design behavioral science experiments with a statistical power of 80% or higher

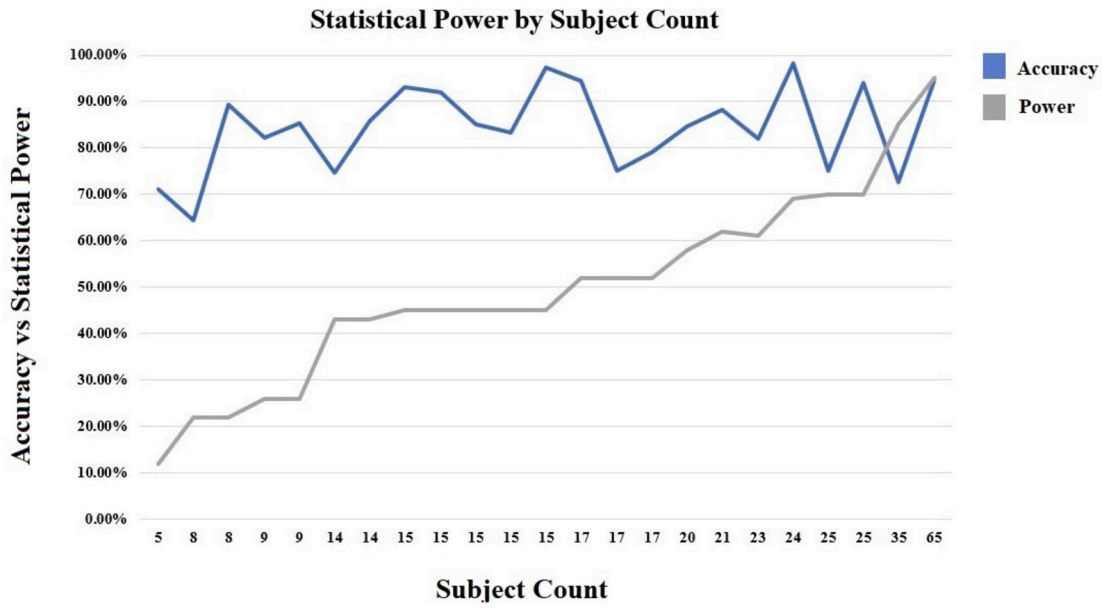

Fig. 6. Accuracy based on number of subjects included in study.

[73], which reduces the probability of encountering a Type II error by up to 20% [84]. Statistical power has three parts: effect size (a statistical measure), sample size (number of observations or participants) and significance (typically 0.05 [73]). Power analysis assists researchers in determining the smallest sample size suitable to detect the effect of a given experiment at a desired level of significance, as collecting larger samples are likely costlier and much harder. The use of machine learning in behavioral science experiments does not automatically negate the need for sufficient statistical power [36,75].

One of the recurrent questions psychology researchers ask is: *"What is the minimum number of participants I must test?"* [74]. The high number of participants required for an 80% powered study often surprises cognitive psychologists, because in their experience, replicable research can be done with a smaller number. For a long time, samples of 20–24 participants were the norm in experimental psychology [74]. However, when applying a two-tailed power test with a correlation coefficient of 0.5 [73] and an assumed significance level of $\alpha = 0.05$, we found that at least 34 test subjects would be required to achieve 80% power. Where correlation is notably less, for example stress biomarker correlation with a periodic stress label, substantially more subjects could be required to achieve at least 80% statistical power. Iqbal et al. [39] specifically performed a power analysis and similarly concluded that at least 34 test subjects would be required to achieve 80% statistical power, and built their Stress-Predict dataset using 35 test subjects.

Considering the small number of subjects contained in the datasets utilized in the experiments reviewed in this paper (Fig. 6), the statistical power of the experiments and subsequent conclusions reached on the accuracy achieved will be overshadowed, more so if these trained models were applied on new, unseen datasets (to confirm generalization). This holds true when the objective is to infer an unknown truth from the observed data, and hypothesis testing provides a specific framework whose inferential target is a binary truth (stressed vs. non-stressed). For example, whether an EDA biomarker from wearable device data provides a signal that correlates with an elevated level of stress. Li et al. [85] provides a detailed discussion and guidelines for choosing between the two strategies (hypothesis testing versus machine learning classification) when designing an experiment that can assist researchers in choosing a strategy and when required, validate whether their sample size contains sufficient power.

4.4. Lack of generalization

Interestingly, as shown in Fig. 6, there appears to be no obvious correlation between the number of subjects included in the study with the reported accuracy rate. In virtually all the reviewed studies the number of subjects were less than 30. None of these studies apart from Mishra et al. [41] and Liapis et al. [47], tested generalization of the resulting models on a totally unseen, new dataset, to further validate the reported accuracy achieved in the experiments, when trained on a public dataset.

Fig. 6 further shows the calculated statistical power given a two-tailed power test with a correlation coefficient of 0.5 [73], and an assumed significance level of $\alpha = 0.05$ for each of the studies reviewed, based on the number of unique test subjects contained within each dataset when used for training and validation. Of these, datasets utilized by Greco et al. [32] and Ehrhart et al. [48] achieved at least 80% power by using non-public datasets while the public WESAD [6,14,27,28,30,45,47] and SWELL [14,49] datasets achieve 45% and 70% power respectively, based on number of test subjects included.

The majority of studies in this review use a custom or public dataset to train their machine learning algorithms using time-series biomarker data within that dataset. These models are then evaluated using the test set of the same dataset, meaning the same experimental setup, and sometimes, different biomarker recordings of previously observed subjects during training. This cannot ensure generalizability of the developed model to other subjects or datasets. Recently some studies are evaluating the potential of person-specific models and their promise in improving generic stress detection models [14]. Of the studies reviews in this paper, and scored using the IJMEDI checklist [2], three studies [32,47,48] were found to likely achieve generalization (Table 8), based on model validation and the use of sufficiently large training datasets based on the number of individual study subjects included.

Focusing specifically on results reported when models are trained and evaluated on the most commonly used WESAD dataset (Table 6), we note that of those experiments reporting accuracy rates higher than 90% [6,14,30,47], all included both EDA and HR (or HRV) biomarkers, while those excluding either the HR or EDA biomarkers [27,28,45] consistently reported accuracy rates below 86%, irrespective of feature-engineering or cross-validation technique applied. This observation is in line with findings by Schmidt et al. [6] who noted their reported highest accuracy rate of 93% dropped to 88.33% when excluding the HR biomarker during model training and validation. This may indi-

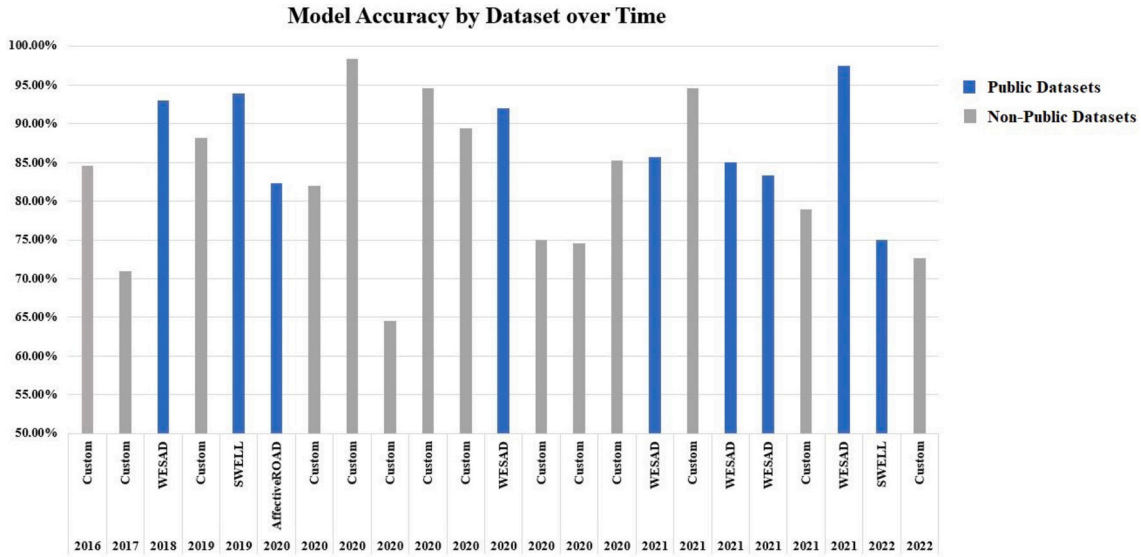

Fig. 7. Reported accuracy based on dataset utilized over time.

cate that both EDA and HR (or derivatives including HRV) biomarkers play an equally important role in correlation with perceived elevated levels of stress and require further examination, considering the small sample of experiments reviewed. Additionally, when considering the advancement of machine learning technologies over the last decade, there appears to be no consistent increase in model performance or reported accuracy over time (Fig. 7), indicating that model generalization with respect to stress detection and measurement using machine learning techniques remains a challenge.

4.5. Summary

The significant observations from this review are:

- Technological improvements in wearable devices have seen a rapid improvement in complexity, ease of use and affordability. This has helped many studies to record and analyze various physiological signals that can be used as biomarkers.
- Sensor biomarkers vary across the wearable devices reviewed, with questions remaining on whether all sensor data can be considered valid and accurate for use in stress detection and measurement, or which biomarkers are the best when measuring stress.
- Existing work has predominantly used small datasets acquired in a single experimental setup with varying labeling protocols, bringing into question the statistical power of these small datasets when used for both training and validation.
- In the studies reviewed, model validation was performed predominantly using LOSO or K-Fold cross-validation, with no further validation on a completely new, unseen dataset recorded in different experimental conditions using new study participants, leaving the question of model generalization unanswered.

4.6. Challenges and future research directions

To achieve reliable machine learning models suitable for real-world monitoring of stress, three formidable challenges should be addressed.

- Varying experimental and labeling protocols influence stress measurement and detection accuracy. To address this challenge, there exists the need for a definitive set of test guidelines when using wearable devices to record biomarker data, including appraisal and scoring methodology. In [86], the authors concluded that, the appraisal process critically shapes an individual's response to acute stress, while [87] detected lower EDA biomarker activity in response to episodes of acute stress in caregivers of people with Autism Spectrum Disorder, a potential habituation to stress. These findings support the need for a proper understanding of when wearable devices can and should be used, and potential factors that could affect sensor accuracy.
- Measurement accuracy is a major challenge that can significantly affect wearable device data and consequently any stress measurements. One of the main problems with current wearables is significant motion artifacts, which may be reduced by measures for better and more stable placement of the device, or through placing the device on other parts of the body.
- Another significant challenge is the lack of large, diverse public datasets built from wearable sensor data that can be utilized to build machine learning models for predicting elevated levels of stress that generalize well to unseen data.

5. Conclusion

The main objective in automated stress detection and measurement is to develop a robust, highly accurate machine learning model that can generalizing well on new, unseen data. The review presented here synthesized the literature and presented important information about the previous studies concerned with stress prediction using wearable devices. In particular, we reviewed and analyzed the publicly available stress biomarker datasets used in numerous studies, the machine learning techniques applied, their advantages, limitations and ability to generalize on new, unseen data. We also summarized our point of view on challenges and opportunities in this emerging domain. We believe this review will advance knowledge in the general area of machine learning for stress detection using wearable devices, helping the research efforts move one step closer to realizing effective stress detection and management technology.

CRediT authorship contribution statement

Gideon Vos, Mostafa Rahimi Azghadi: Conceptualization, Data Analysis and Interpretation. **Gideon Vos**: Writing-Original draft preparation, Investigation, Visualization. **Mostafa Rahimi Azghadi**: Writing-Critical Revision, Reviewing, and Editing. **Kelly Trinh, Zoltan Sarnyai**: Writing-Reviewing and Editing. **Kelly Trinh, Zoltan Sarnyai, Mostafa Rahimi Azghadi**: Supervision.

Declaration of competing interest

The authors, Gideon Vos, Kelly Trinh, Zoltan Sarnyai, and Mostafa Rahimi Azghadi, declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2023.105026.

References

[1] P. Boucher, P. Plusquellec, Acute stress assessment from excess cortisol secretion: fundamentals and perspectives, Front. Neuroendocrinol. 10 (nov 2019), https://doi.org/10.3389/fendo.2019.00749.

[2] F. Cabitza, A. Campagner, The need to separate the wheat from the chaff in medical informatics, Int. J. Med. Inform. 153 (2021) 104510, https://doi.org/10.1016/j.ijmedinf.2021.104510.

[3] B.S. McEwen, Protective and damaging effects of stress mediators, N. Engl. J. Med. 338 (3) (1998) 171–179, https://doi.org/10.1056/nejm199801153380307.

[4] 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 Invest. 15 (3) (2018) 235–245, https://doi.org/10.30773/pi.2017.08.17.

[5] J.F. Thayer, F. Åhs, M. Fredrikson, J.J. Sollers, T.D. Wager, A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health, Neurosci. Biobehav. Rev. 36 (2) (2012) 747–756, https://doi.org/10.1016/j.neubiorev.2011.11.009.

[6] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, K.V. Laerhoven, Introducing WESAD, a multimodal dataset for wearable stress and affect detection, in: Proceedings of the 20th ACM International Conference on Multimodal Interaction, ACM, 2018.

[7] W. Kraaij, S. Koldijk, M. Sappelli, The swell knowledge work dataset for stress and user modeling research, https://doi.org/10.17026/DANS-X55-69ZP, 2015.

[8] A. Rossi, E. Da Pozzo, D. Menicagli, C. Tremolanti, C. Priami, A. Sirbu, D. Clifton, C. Martini, D. Morelli, Multilevel monitoring of activity and sleep in healthy people, https://doi.org/10.13026/CERQ-FC86, 2020.

[9] P.J. Brantley, C.D. Waggoner, G.N. Jones, N.B. Rappaport, A daily stress inventory: development, reliability, and validity, J. Behav. Med. 10 (1) (1987) 61–73, https://doi.org/10.1007/bf00845128.

[10] N.E. Haouij, J.-M. Poggi, S. Sevestre-Ghalila, R. Ghozi, M. Jaïdane, AffectiveROAD system and database to assess driver's attention, in: Proceedings of the 33rd Annual ACM Symposium on Applied Computing, ACM, 2018.

[11] P. Siirtola, J. Röning, Comparison of regression and classification models for user-independent and personal stress detection, Sensors 20 (16) (2020) 4402, https://doi.org/10.3390/s20164402.

[12] Y.S. Can, D. Gokay, D.R. Kılıç, D. Ekiz, N. Chalabianloo, C. Ersoy, How laboratory experiments can be exploited for monitoring stress in the wild: a bridge between laboratory and daily life, Sensors 20 (3) (2020) 838, https://doi.org/10.3390/s20030838.

[13] M. Gjoreski, M. Luštrek, M. Gams, H. Gjoreski, Monitoring stress with a wrist device using context, J. Biomed. Inform. 73 (2017) 159–170, https://doi.org/10.1016/j.jbi.2017.08.006.

[14] K. Nkurikiyeyezu, A. Yokokubo, G. Lopez, The effect of person-specific biometrics in improving generic stress predictive models, arXiv:1910.01770, Oct. 2019.

[15] Fitbit.com, Fitbit, http://www.fitbit.com, 1 2022.

[16] Empatica, Empatica | medical devices, ai and algorithms for remote patient monitoring, http://www.empatica.com, 1 2022.

[17] J.M. Peake, G. Kerr, J.P. Sullivan, A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations, Front. Physiol. 9 (jun 2018), https://doi.org/10.3389/fphys.2018.00743.

[18] G. Regalia, F. Onorati, M. Lai, C. Caborni, R.W. Picard, Multimodal wrist-worn devices for seizure detection and advancing research: focus on the empatica wristbands, Epilepsy Res. 153 (2019) 79–82, https://doi.org/10.1016/j.eplepsyres.2019.02.007.

[19] J. Tang, R.E. Atrache, S. Yu, U. Asif, M. Jackson, S. Roy, M. Mirmomeni, S. Cantley, T. Sheehan, S. Schubach, C. Ufongene, S. Vieluf, C. Meisel, S. Harrer, T. Loddenkemper, Seizure detection using wearable sensors and machine learning: setting a benchmark, Epilepsia 62 (8) (2021) 1807–1819, https://doi.org/10.1111/epi.16967.

[20] F. Onorati, G. Regalia, C. Caborni, W.C. LaFrance, A.S. Blum, J. Bidwell, P.D. Liso, R.E. Atrache, T. Loddenkemper, F. Mohammadpour-Touserkani, R.A. Sarkis, D. Friedman, J. Jeschke, R. Picard, Prospective study of a multimodal convulsive seizure detection wearable system on pediatric and adult patients in the epilepsy monitoring unit, Front. Neurol. 12 (aug 2021), https://doi.org/10.3389/fneur.2021.724904.

[21] N. Kulkarni, C. Compton, J. Luna, M.A.U. Alam, Monitoring my dehydration: a non-invasive dehydration alert system using electrodermal activity, arXiv:2009.13626, Sep. 2020.

[22] M. Gjoreski, T. Kolenik, T. Knez, M. Luštrek, M. Gams, H. Gjoreski, V. Pejović, Datasets for cognitive load inference using wearable sensors and psychological traits, Appl. Sci. 10 (11) (2020) 3843, https://doi.org/10.3390/app10113843.

[23] D. Gholamiangonabadi, N. Kiselov, K. Grolinger, Deep neural networks for human activity recognition with wearable sensors: leave-one-subject-out cross-validation for model selection, IEEE Access 8 (2020) 133982–133994, https://doi.org/10.1109/access.2020.3010715.

[24] T. Zhang, A.E. Ali, C. Wang, A. Hanjalic, P. Cesar, CorrNet: fine-grained emotion recognition for video watching using wearable physiological sensors 21 (1) (2020) 52, https://doi.org/10.3390/s21010052.

[25] E.E. Kaczor, S. Carreiro, J. Stapp, B. Chapman, P. Indic, Objective measurement of physician stress in the emergency department using a wearable sensor, in: Proceedings of the Annual Hawaii International Conference on System Sciences 2020, 2020, pp. 3729–3738.

[26] E. Smets, P. Casale, U. Großekathöfer, B. Lamichhane, W.D. Raedt, K. Bogaerts, I.V. Diest, C.V. Hoof, Comparison of Machine Learning Techniques for Psychophysiological Stress Detection, Springer International Publishing, 2016, pp. 13–22.

[27] M. Alshamrani, An advanced stress detection approach based on processing data from wearable wrist devices, Int. J. Adv. Comput. Sci. Appl. 12 (7) (2021), https://doi.org/10.14569/ijacsa.2021.0120745.

[28] T. Iqbal, P. Redon-Lurbe, A.J. Simpkin, A. Elahi, S. Ganly, W. Wijns, A. Shahzad, A sensitivity analysis of biophysiological responses of stress for wearable sensors in connected health, IEEE Access 9 (2021) 93567–93579, https://doi.org/10.1109/access.2021.3082423.

[29] Y.S. Can, N. Chalabianloo, D. Ekiz, C. Ersoy, Continuous stress detection using wearable sensors in real life: algorithmic programming contest case study, Sensors 19 (8) (2019) 1849, https://doi.org/10.3390/s19081849.

[30] F.I. Indikawati, S. Winiarti, Stress detection from multimodal wearable sensor data, IOP Conf. Ser., Mater. Sci. Eng. 771 (1) (2020) 012028, https://doi.org/10.1088/1757-899x/771/1/012028.

[31] H.J. Han, S. Labbaf, J.L. Borelli, N. Dutt, A.M. Rahmani, Objective stress monitoring based on wearable sensors in everyday settings, J. Med. Eng. Technol. 44 (4) (2020) 177–189, https://doi.org/10.1080/03091902.2020.1759707.

[32] A. Greco, G. Valenza, J. Lazaro, J.M. Garzon-Rey, J. Aguilo, C.D. la Camara, R. Bailon, E.P. Scilingo, Acute stress state classification based on electrodermal activity modeling, IEEE Trans. Affect. Comput. (2021), https://doi.org/10.1109/taffc.2021.3055294.

[33] M. Sevil, M. Rashid, M.R. Askari, Z. Maloney, I. Hajizadeh, A. Cinar, Detection and characterization of physical activity and psychological stress from wristband data, Signals 1 (2) (2020) 188–208, https://doi.org/10.3390/signals1020011.

[34] C. Samson, A. Koh, Stress monitoring and recent advancements in wearable biosensors, Front. Bioeng. Biotechnol. 8 (sep 2020), https://doi.org/10.3389/fbioe.2020.01037.

[35] S. Gedam, S. Paul, A review on mental stress detection using wearable sensors and machine learning techniques, IEEE Access 9 (2021) 84045–84066, https://doi.org/10.1109/access.2021.3085502.

[36] R.D. Riley, J. Ensor, K.I.E. Snell, F.E. Harrell, G.P. Martin, J.B. Reitsma, K.G.M. Moons, G. Collins, M. van Smeden, Calculating the sample size required for developing a clinical prediction model, BMJ (2020), https://doi.org/10.1136/bmj.m441.

[37] C.Y. Park, Narae Cha, Soowon Kang, A. Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, A. Oh, Y. Jeong, Uichin Lee, K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations, https://doi.org/10.5281/ZENODO.3931963, 2020.

[38] H. Svoren, V. Thambawita, P. Halvorsen, P. Jakobsen, E.G. Ceja, F.M. Noori, H.L. Hammer, M. Lux, M. Riegler, S. Hicks, Toadstool: a dataset for training emotional intelligent machines playing super mario bros, https://doi.org/10.31219/osf.io/4v9mp, feb 2020.

[39] T. Iqbal, A.J. Simpkin, D. Roshan, N. Glynn, J. Killilea, J. Walsh, G. Molloy, S. Ganly, H. Ryman, E. Coen, A. Elahi, W. Wijns, A. Shahzad, Stress monitoring using wearable sensors: a pilot study and stress-predict dataset, Sensors 22 (21) (2022) 8135, https://doi.org/10.3390/s22218135.

[40] E. Smets, E.R. Velazquez, G. Schiavone, I. Chakroun, E. D'Hondt, W.D. Raedt, J. Cornelis, O. Janssens, S.V. Hoecke, S. Claes, I.V. Diest, C.V. Hoof, Large-scale wearable data reveal digital phenotypes for daily-life stress detection, npj Digit. Med. 1 (1) (dec 2018), https://doi.org/10.1038/s41746-018-0074-9.

[41] 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, https://doi.org/10.1145/3432220.

[42] C.W. Jin, A. Osotsi, Z. Oravecz, Predicting stress in teens from wearable device data using machine learning methods, https://doi.org/10.1101/2020.11.26.20223784, dec 2020.

[43] F. Delmastro, F.D. Martino, C. Dolciotti, Cognitive training and stress detection in MCI frail older people through wearable sensors and machine learning, IEEE Access 8 (2020) 65573–65590, https://doi.org/10.1109/access.2020.2985301.

[44] M. Elgendi, C. Menon, Machine learning ranks ecg as an optimal wearable biosignal for assessing driving stress, IEEE Access 8 (2020) 34362–34374, https://doi.org/10.1109/access.2020.2974933.

[45] P. Garg, J. Santhosh, A. Dengel, S. Ishimaru, Stress detection by machine learning and wearable sensors, https://doi.org/10.1145/3397482.3450732, 2021.
[46] K.M. Dalmeida, G.L. Masala, HRV features as viable physiological markers for stress detection using wearable devices, Sensors 21 (8) (2021) 2873, https://doi.org/10.3390/s21082873.
[47] A. Liapis, E. Faliagka, C.P. Antonopoulos, G. Keramidas, N. Voros, Advancing stress detection methodology with deep learning techniques targeting UX evaluation in AAL scenarios: applying embeddings for categorical variables, Electronics 10 (13) (2021) 1550, https://doi.org/10.3390/electronics10131550.
[48] M. Ehrhart, B. Resch, C. Havas, D. Niederseer, A conditional GAN for generating time series data for stress detection in wearable physiological sensor data, Sensors 22 (16) (2022) 5969, https://doi.org/10.3390/s22165969.
[49] T. Iqbal, A. Elahi, W. Wijns, A. Shahzad, Exploring unsupervised machine learning classification methods for physiological stress detection, Front. Med. Technol. 4 (mar 2022), https://doi.org/10.3389/fmedt.2022.782756.
[50] M. Swan, The quantified self: fundamental disruption in big data science and biological discovery, Big Data 1 (2) (2013) 85–99, https://doi.org/10.1089/big.2012.0002.
[51] J.H.D.M. Westerink, R.J.E. Rajae-Joordens, M. Ouwerkerk, M. van Dooren, S. Jelfs, A.J.M. Denissen, E.P. de Vries, R. van Ee, Deriving a cortisol-related stress indicator from wearable skin conductance measurements: quantitative model & experimental validation, Front. Comput. Sci. 2 (sep 2020), https://doi.org/10.3389/fcomp.2020.00039.
[52] M.R. Azghadi, C. Lammie, J.K. Eshraghian, M. Payvand, E. Donati, B. Linares-Barranco, G. Indiveri, Hardware implementation of deep network accelerators towards healthcare and biomedical applications 14 (6) (2020) 1138–1159, https://doi.org/10.1109/tbcas.2020.3036081.
[53] P. Siirtola, Continuous stress detection using the sensors of commercial smartwatch, https://doi.org/10.1145/3341162.3344831, 2019, pp. 1198–1201.
[54] T.F.D. Farrow, N.K. Johnson, M.D. Hunter, A.T. Barker, I.D. Wilkinson, P.W.R. Woodruff, Neural correlates of the behavioral-autonomic interaction response to potentially threatening stimuli, Front. Human Neurosci. 6 (2013), https://doi.org/10.3389/fnhum.2012.00349.
[55] M. Saifuzzaman, T.N. Ananna, M.J.M. Chowdhury, M.S. Ferdous, F. Chowdhury, A systematic literature review on wearable health data publishing under differential privacy, Int. J. Inf. Secur. 21 (4) (2022) 847–872, https://doi.org/10.1007/s10207-021-00576-1.
[56] R. Simons, R. Koordeman, P. de Looff, R. Otten, Physiological measurements of stress preceding incidents of challenging behavior in people with severe to profound intellectual disabilities: longitudinal study protocol of single-case studies, JMIR Res. Protoc. 10 (7) (2021) e24911, https://doi.org/10.2196/24911.
[57] K. Albrecht, Stress and the Manager: Making It Work for You, 1979.
[58] A. Ali, S.M. Shamsuddin, A. Ralescu, Classification with class imbalance problem: a review 7 (2015) 176–204.
[59] G. Deng, C. Han, T. Dreossi, C. Lee, D.S. Matteson, IB-GAN: a unified approach for multivariate time series classification under class imbalance, in: Proceedings of the 2022 SIAM International Conference on Data Mining (SDM), Society for Industrial and Applied Mathematics, 2022, pp. 217–225.
[60] W. Lee, K. Seo, Downsampling for binary classification with a highly imbalanced dataset using active learning, Big Data Res. 28 (2022) 100314, https://doi.org/10.1016/j.bdr.2022.100314.
[61] Z. Jiang, T. Pan, C. Zhang, J. Yang, A new oversampling method based on the classification contribution degree, Symmetry 13 (2) (2021) 194, https://doi.org/10.3390/sym13020194.
[62] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357, https://doi.org/10.1613/jair.953.
[63] H. He, Y. Bai, E.A. Garcia, S. Li, ADASYN: adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IEEE, 2008.
[64] J. Brandt, E. Lanzen, A comparative review of smote and adasyn in imbalanced data classification, 2021.
[65] J.L. Leevy, T.M. Khoshgoftaar, R.A. Bauder, N. Seliya, A survey on addressing high-class imbalance in big data, J. Big Data 5 (1) (nov 2018), https://doi.org/10.1186/s40537-018-0151-6.
[66] D. Elreedy, A.F. Atiya, F. Kamalov, A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning, Mach. Learn. (jan 2023), https://doi.org/10.1007/s10994-022-06296-4.
[67] I. Pastaltzidis, N. Dimitriou, K. Quezada-Tavarez, S. Aidinlis, T. Marquenie, A. Gurzawska, D. Tzovaras, Data augmentation for fairness-aware machine learning, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, ACM, 2022.
[68] C. Guo, M. Lu, J. Chen, An evaluation of time series summary statistics as features for clinical prediction tasks, BMC Med. Inform. Decis. Mak. 20 (1) (mar 2020), https://doi.org/10.1186/s12911-020-1063-x.
[69] S. Kreibig, Autonomic nervous system activity in emotion: a review, Biol. Psychol. 84 (2010) 394–421, https://doi.org/10.1016/j.biopsycho.2010.03.010.
[70] E. Elgeldawi, A. Sayed, A.R. Galal, A.M. Zaki, Hyperparameter tuning for machine learning algorithms used for Arabic sentiment analysis, Informatics 8 (4) (2021), https://doi.org/10.3390/informatics8040079, https://www.mdpi.com/2227-9709/8/4/79.
[71] T. Umematsu, A. Sano, S. Taylor, M. Tsujikawa, R.W. Picard, Forecasting stress, mood, and health from daytime physiology in office workers and students, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2020.
[72] T. Iqbal, A. Elahi, P. Redon, P. Vazquez, W. Wijns, A. Shahzad, A review of biophysiological and biochemical indicators of stress for connected and preventive healthcare, Diagnostics 11 (3) (2021) 556, https://doi.org/10.3390/diagnostics11030556.
[73] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum Associates, 1988.
[74] M. Brysbaert, How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables 2 (1), https://doi.org/10.5334/joc.72, 2019.
[75] Sample size in machine learning and artificial intelligence - perioperative data science, 10 2022.
[76] H.G. van Lier, M.E. Pieterse, A. Garde, M.G. Postel, H.A. de Haan, M.M.R. Vollenbroek-Hutten, J.M. Schraagen, M.L. Noordzij, A standardized validity assessment protocol for physiological signals from wearable technology: methodological underpinnings and an application to the e4 biosensor, Behav. Res. Methods 52 (2) (2019) 607–629, https://doi.org/10.3758/s13428-019-01263-9.
[77] N. Milstein, I. Gordon, Validating measures of electrodermal activity and heart rate variability derived from the empatica e4 utilized in research settings that involve interactive dyadic states, Front. Behav. Neurosci. 14 (aug 2020), https://doi.org/10.3389/fnbeh.2020.00148.
[78] W. Ryan, J. Conigrave, G. Basarkod, J. Ciarrochi, B.K. Sahdra, When is it good to use wristband devices to measure HRV?: introducing a new method for evaluating the quality of data from photophlethysmography-based HRV devices, https://doi.org/10.31234/osf.io/t3gdz, may 2019.
[79] K. Georgiou, A.V. Larentzakis, N.N. Khamis, G.I. Alsuhaibani, Y.A. Alaska, E.J. Giallafos, Can wearable devices accurately measure heart rate variability? A systematic review, Folia Medica 60 (1) (jan 2018), https://doi.org/10.2478/folmed-2018-0012.
[80] A.A.T. Schuurmans, P. de Looff, K.S. Nijhof, C. Rosada, R.H.J. Scholte, A. Popma, R. Otten, Validity of the empatica e4 wristband to measure heart rate variability (HRV) parameters: a comparison to electrocardiography (ECG), J. Med. Syst. 44 (11) (sep 2020), https://doi.org/10.1007/s10916-020-01648-w.
[81] S.M. Seipäjärvi, A. Tuomola, J. Juurakko, M. Rottensteiner, A.-P.E. Rissanen, J.L.O. Kurkela, U.M. Kujala, J.A. Laukkanen, J. Wikgren, Measuring psychosocial stress with heart rate variability-based methods in different health and age groups, Physiol. Meas. 43 (5) (2022) 055002, https://doi.org/10.1088/1361-6579/ac6b7c.
[82] P. Schmidt, R. Dürichen, A. Reiss, K.V. Laerhoven, T. Plötz, Multi-target affect detection in the wild, in: Proceedings of the 23rd International Symposium on Wearable Computers, ACM, 2019.
[83] S. Ghiasi, A. Greco, R. Barbieri, E.P. Scilingo, G. Valenza, Assessing autonomic function from electrodermal activity and heart rate variability during cold-pressor test and emotional challenge, Sci. Rep. 10 (1) (mar 2020), https://doi.org/10.1038/s41598-020-62225-2.
[84] J. Brownlee, A gentle introduction to statistical power and power analysis in Python, Apr 2020.
[85] J.J. Li, X. Tong, Statistical hypothesis testing versus machine learning binary classification: distinctions and guidelines, Patterns 1 (7) (2020) 100115, https://doi.org/10.1016/j.patter.2020.100115.
[86] E.S. Epel, A.D. Crosswell, S.E. Mayer, A.A. Prather, G.M. Slavich, E. Puterman, W.B. Mendes, More than a feeling: a unified view of stress measurement for population science, Front. Neuroendocrinol. 49 (2018) 146–169, https://doi.org/10.1016/j.yfrne.2018.03.001.
[87] N. Ruiz-Robledillo, L. Moya-Albiol, Lower electrodermal activity to acute stress in caregivers of people with autism spectrum disorder: an adaptive habituation to stress, J. Autism Dev. Disord. 45 (2) (2013) 576–588, https://doi.org/10.1007/s10803-013-1996-3.