Graphical Abstract

Machine Learning for Stress Monitoring from Wearable Devices: A Systematic Literature Review

Gideon Vos, Kelly Trinh, Zoltan Sarnyai, Mostafa Rahimi Azghadi
Machine learning is increasingly used for health monitoring using wearable device sensor data, including the measurement and detection of elevated levels of stress.

We reviewed the literature on using machine learning for stress detection, with an emphasis on their potential to generalize on new, unseen data.

While significant advances have been made, more research is needed to build large, varied datasets for training machine learning models capable of generalizing on new, unseen data and experimental conditions.
Highlights

Machine Learning for Stress Monitoring from Wearable Devices: A Systematic Literature Review

Gideon Vos, Kelly Trinh, Zoltan Sarnyai, Mostafa Rahimi Azghadi

- Wearable sensors for health monitoring have become rapidly available and more sophisticated since 2009, and there is an increased interest in applying machine learning techniques to wearable sensors data for stress monitoring.

- This paper provides a review of the current state of stress detection and measurement from wearable devices using machine learning. The reviewed works are synthesized into three categories of publicly available stress datasets, machine learning, and future research directions. We also review wearable devices with a focus on those capable of recording four important stress biomarkers.

- Our review of machine learning models is provided by analyzing and synthesizing the literature based on seven critical development steps in machine learning pipeline.

- We show that most stress-related machine learning studies are performed on small, singular datasets with a lack of generalization, and larger studies that combine or build substantially more varied datasets are needed.

- We provide a critical review on the shortcomings of previous works such as their labeling protocols, lack of statistical power, validity of stress biomarkers, and lack of generalization.

- We highlight a few of the prominent challenges and propose future research opportunities in the area of machine learning for stress detection using wearables.
Machine Learning for Stress Monitoring from Wearable Devices: A Systematic Literature Review

Gideon Vos\textsuperscript{a}, Kelly Trinh\textsuperscript{a}, Zoltan Sarnyai\textsuperscript{b}, Mostafa Rahimi Azghadi\textsuperscript{a}

\textsuperscript{a}College of Science and Engineering, James Cook University, James Cook Dr, Townsville, 4811, QLD, Australia
\textsuperscript{b}College of Public Health, Medical, and Vet Sciences, James Cook University, James Cook Dr, Townsville, 4811, QLD, Australia

Abstract

Introduction. Wearable sensors have shown promise as a non-intrusive method for collecting biomarkers that may correlate with levels of elevated stress. The stress response has both subjective, psychological and objectively measurable, biological components. Both of them can be expressed differently from person to person, complicating the development of a generic stress measurement model. This is further compounded by the lack of large, labeled datasets that can be utilized to build machine learning models for accurately detecting periods and levels of stress. The datasets publicly available are usually collected using different devices and experimental settings, and are labeled using differing scoring methods. The aim of this review is to provide an overview of the current state of stress detection and monitoring using wearable devices, and where applicable, machine learning techniques utilized. 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 datasets designed for detecting stress and their associated machine learning methods, with a systematic review and meta-analysis of those that utilized wearable sensor data as stress biomarkers. The electronic databases of Google Scholar, Crossref, DOAJ and PubMed were searched for relevant articles and a total of 24 articles were identified and included in the final analysis. The reviewed works were synthesized into three categories of publicly available stress datasets, machine learning, and future research directions. For the machine learning studies reviewed, we provide an analysis of their approach.
to the different steps of a machine learning development pipeline, further synthesizing the finding and trends in the literature. We also provide a review of several major wearable devices that are able to measure four biomarkers known to be robust indicators of elevated levels of stress.

**Results.** A wide variety of study-specific test and measurement protocols were noted in the literature. A number of public datasets were identified that are labeled for stress detection. We focused mainly on datasets using the Empatica E4 device, a well-studied, easy to access, United States Food and Drug Administration (FDA) approved wrist-worn wearable that provides sensor biomarkers most notable to correlate with elevated levels of stress. Most 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 generalization ability.

**Conclusion.** Health tracking and monitoring using wearable devices is growing in popularity, and continued interest with a growing number of studies in this area of health care will likely improve the accuracy of detecting specific health conditions including elevated levels of stress. There is a need for a definitive device-use and setup protocol to enable consistent results across tests with medical-grade wearable devices. Generalization of existing machine learning models still require further study, and research in this area will continue to provide improvements as newer and more substantial datasets become available for study.

**Keywords:** Stress, Wearable sensor, Empatica E4, Machine learning

**PACS:** 07.05.Mh, 87.85.fk

**2000 MSC:** 68T01, 92C99

---

1. **Introduction**

Wearable devices for personal health monitoring and tracking have gained significant popularity and technical sophistication since the release of the first Fitbit [1] in 2009 and Empatica Embrace model in 2016 [2]. Recently, more advanced devices including Empatica’s E4 [2] have been developed that are capable of measuring a wide variety of physiological signals. Continuous measurement of these signals using wearables enable researchers to extract
useful information from these devices to detect and monitor a variety of potential health-related events such as seizures [3–5], dehydration [6], cognitive load [7], physical activity [8], emotions [9] and specifically related to this review, stress [7, 10–21].

The biophysical and biochemical markers normally used to measure stress include Electrodermal Activity (EDA), Electrocardiography (ECG), Electroencephalograph (EEG), Respiration Rate (RR), Blood Pressure (BP), Blood Volume Pulse (BVP), Skin temperature (ST), Electromyography (EMG), plasma catecholamines, copeptin and prolactin, steroid samples, α-amylase and cortisol samples [22]. A wearable device needs to be able to record at least some of these biomarkers, or a derivative thereof such as Heart Rate (HR), in order to detect and measure the biological effects of stress.

Despite some limitations such as battery time and incorrect placement [16], compared to controlled laboratory measurement devices, wearables are non-intrusive and very easy to use. This ease has facilitated many experiments using wearables, and predominantly utilizing Empatica’s latest E4 device, which have yielded a number of well-studied public datasets [23–27]. Consequently, further studies have been conducted to build machine learning models capable of detecting elevated levels of stress, within these datasets [7, 10–21].

A number of previous survey articles have studied the topics of stress detection using wearable devices [28] and machine learning [29]. In particular, in [28], 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 [29], 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 the reliable medical-grade Empatica devices and biomarkers different to those used in [29]. We believe this makes our analysis more realistic.

In addition, the previous reviews have not addressed a number of important
points such as the statistical power of utilized datasets and their labeling protocols and how it affects machine learning techniques, or validity of different sensor biomarkers for stress detection. Also, they have not provided a synthesize of the state-of-the-art machine learning works and how they approach different steps of machine learning pipelines to develop effective models. Another essential question not covered is whether machine learning techniques built on any of the previous datasets can yield a model 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.

This paper is motivated by this question, as developing many separate stress-related wearable datasets, and tailored machine learning techniques for them, will not be very effective in reaching generalizable stress detection methods for wearable devices. Towards addressing this question, we explore the current state of stress detection and measurement using reliable state-of-the-art devices, with an emphasis on using machine learning to automate the measurement analysis. We further explore the available public datasets built using recorded sensor data from the medical-grade Empatica E4 device, and investigate the modeling approaches utilized and detection accuracy scores attained for machine learning models trained on these datasets. Finally, we discuss the feasibility of using these machine learning models and associated public datasets to measure stress on unseen data when deployed against new experimental results, in order to understand the current state of using wearable devices for accurately measuring stress response.

In summary, this paper covers the following aspects of stress detection using machine learning applied to wearable device data: (i) The use of wearable devices for data recording and detecting elevated levels of stress. (ii) The availability of reliable public datasets to utilize for training new machine learning models capable of detecting and measuring elevated levels of stress. (iii) A synthesis and a critical analysis of machine learning techniques previously utilized to successfully detect elevated levels of stress. (iv) The generalization capabilities of the various techniques reviewed and where applicable, the datasets utilized for model training, and (v) the future direction and open research questions to further the main scientific objective of accurately detecting and measuring elevated levels of stress using wearable devices.
The remainder of this paper is organized in the following sections: Section 2 provides a brief definition of stress, the measurement thereof, and how machine learning is being utilized to measure stress; Section 3 describes review focus and methods; Section 4 highlights prominent results and limitations; and Section 5 provides a detailed discussion and future directions. Finally, Section 6 briefly concludes this study.

2. Definitions

2.1. Stress

Stress can be defined as a challenge to an individual (human and animal) that overtaxes their control systems or exceeds their personal resources, disrupting homeostasis and resulting in adverse effects. In response to a stressful stimulus, a coordinated biological response first originates in the brain and then spreads to the whole body. Cortical centers in the brain sense stress and activate the autonomic nervous system (ANS) and the hypothalamic-pituitary-adrenal (HPA) axis. The ANS is a component of the peripheral nervous system that regulates involuntary physiological processes including heart rate, blood pressure, respiration, digestion, and sexual arousal. It contains three anatomically distinct divisions: sympathetic, parasympathetic, and enteric [30]. During stress, the sympathetic nervous system (SNS) activity is predominant, with the parasympathetic system providing a counterbalance.

Activation of the SNS leads to a state of overall elevated activity and attention: the “fight or flight” response. The ANS quickly promotes physiological changes through the SNS and parasympathetic nervous system (PNS). During the stress response, the HPA axis triggers a series of endocrine changes, characterized by the release of corticotropin-releasing hormone (CRH) from the hypothalamus, which in turn activates the synthesis and release of the adrenocorticotropic hormone (ACTH) from the anterior pituitary. ACTH then reaches the outer layers of the adrenal gland (adrenal cortex) through the systemic circulation to release cortisol. The PNS plays an important role in alleviating the stress response of individuals by inhibiting the SNS and HPA axis. The ANS responds to the needs of the internal viscera as well as external stimuli. Homeostasis is associated with the regulation of internal viscera, whereas the stress response prioritizes external stimuli over internal
Situations that activate the HPA axis as part of a brief emergency response, but whose effects are useful to the individual, would not generally be considered as a stressor [31]. The biological and behavioral stress response is adaptive in the short term. They help to mobilize the organism to deal with the stressful stimulus. However, if stress becomes chronic the powerful bio-molecules, such as adrenalin, released by the activation of the SNS, and cortisol, elevated through the activation of the HPA axis, will have long-term negative consequences on body and brain health. Research from the past decade illustrates that sympathetic and parasympathetic interactions are more complex than previously assumed. Patterns of ANS activation vary across individuals, with distinct physiological response profiles influencing the reactivity underlying automatic behavioral responses [32]. These factors make stress measurement and monitoring at an individual level challenging.

2.2. Stress Measurement

Stress responses can be measured with self-report measures, behavioral coding, or via physiological measurements. These responses include emotions, cognition, behaviors, and physiological responses instigated by the stressful stimuli. One of the simplest ways to measure stress responses is through self-reports of perceived stress related to a specific stressor or to one’s life circumstances [33].

Acute (episodic) stressors have relatively discreet beginnings and endings, whereas chronic stress can be viewed as the emotional and physiological impacts experienced over a prolonged period of time, during which the individual feels they have little or no control [34]. Stress detection and measurement is an active area of research [35] with important implications for the personal, professional, and the overall health of an individual. Chronic stress is highly detrimental to physiological health and psychological well-being [36], and developing robust, non-intrusive methods for the accurate detection and measurement of stress is of great importance to healthcare service providers.
2.3. Stress Measurement Using Wearable Sensors and Machine Learning

Currently, stress-related and personalized questionnaires are mainly used to measure or score (label) stress in real-life and outside of a laboratory context. However, this technique does not allow for continuous monitoring, and often suffers from bias such as demand effects, response and memory biases. Therefore, the focus has shifted towards measuring bodily responses as indicators of stress [13]. However, these are not ideal since the measurement techniques are often intrusive and difficult to measure continuously. Technology offers wearable devices as a solution that allows researchers to gather and utilize large quantities of sensor-based physiological data continuously and non-intrusively, and automate the process of exploration, analysis and measurement through the use of machine learning techniques.

Machine learning is a computer-based system (see Figure 1) that can learn and adapt without following explicit instructions, through the use of algorithms and statistical models. Machine learning is divided into supervised, unsupervised, semi-supervised, and reinforcement learning [29]. From the literature reviewed, supervised and unsupervised techniques are commonly utilized for predicting elevated levels of stress. In supervised learning, models are trained using data that is accurately labeled for the response you are predicting for; in the context of this paper, the labeling would be for elevated levels of stress with labels as binary yes/no indicators or a numeric scale to indicate stress level, generally a range between 0 (no stress) and 1 (maximum stress). Unsupervised learning, in contrast, does not require labeling, but instead allows the model to discover information on its own. When dealing with supervised learning problems, accurate labeling is critical to ensure proper learning occurs within the model.

The goal of a good machine learning model is to generalize well from the training data to any new data from within the problem domain. Generalization refers to how well a trained machine learning model can perform on unseen data, i.e., data not included when initially training the model. A generalized machine learning model, therefore, implies a model that can predict on new, unseen data as well as it performed on the training data. In order to generate such a model, the training dataset should be diverse. Within the context of this paper, this implies recording data samples under varying experimental conditions, and across demographics. Variance in the context of machine learning relates to the variety of predictions made by the model,
while Bias refers to the distance of the predictions from the actual (true) values. A highly-biased model implies its predicted values are far from the actual, true values.

A generalized model offers the best trade-off between bias and variance, thereby delivering the best predictive performance. In [18, 37], the authors argue that many of the published physiological-based machine learning stress detection models may not be practical in real-world settings, due to problems with generalization. Nkurikiyeyezu et al. [18] have proposed a method that shows the possibility of designing a stress recognition system that is based on generic stress recognition models, but can be tuned by incorporating the physiological fingerprints of new, unseen subjects. Furthermore, [38] has noted that stress response is person-specific, and may require training models specifically to their (person-specific) data for accurate stress prediction. This review is written considering these critical points to help achieve improvements in more generalizable machine learning models for detecting stress from wearable device data.
3. Methods

3.1. Research questions

The main aim of this work is to provide an overview of the current state of stress detection using wearable devices, and where applicable, machine learning techniques utilized, their findings and results, and ability of their models to generalize on new, unseen data. Thus, our research questions can be formulated as follows:

- **RQ1**: What is the current state of stress detection using wearable devices?
- **RQ2**: What are the most successful machine learning techniques utilized?

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.

### Table 1: Queries used for our search, expressed in pseudo-code.

| Query | Description |
|-------|-------------|
| Q1    | *in title or abstract*: stress |
| Q2    | *in title or abstract*: stress AND (data AND ((wearable OR empatica) OR ("machine learning" OR "artificial intelligence"))) |

3.2. Search strategy

We reviewed key published works between 2012 and 2021 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 JabRef reference management software using the queries detailed in Table 1 and a total of 788 papers were identified with the first query.

A subsequent query was run on these results for further filtering, resulting in 105 papers as shown in Figure 2. JabRef was then used to identify duplicates and 16 were found and removed, leaving the number of considered
papers for the subsequent phases at 89. 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 animal or psychiatry, were excluded. Studies using devices that are generally considered as health-trackers, or lifestyle monitors were also excluded, as was studies performed solely using devices that would not generally be considered a wearable device, such as EEG or chest-worn monitors. Furthermore, public datasets predominantly built with data recorded on those devices were excluded, and studies conducted on datasets containing less than 5 subjects were also excluded. Relevant papers were considered based on title, keywords and abstracts where the focus was predominantly on stress, and more specifically stress detection and measurement using wearable devices where machine learning techniques were applied.

As a result, a total of 24 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 2 details the papers included in this review.
Table 2: Studies included in this review.

| Topic                | Reference | Paper                                                                 | Date  |
|----------------------|-----------|-----------------------------------------------------------------------|-------|
| Datasets             | [39]      | The swell knowledge work dataset for stress and user modeling research| 2015  |
| Datasets             | [23]      | Introducing WESAD, a multimodal dataset for wearable stress and affect detection | 2018  |
| Datasets             | [25]      | AffectiveROAD system and database to assess driver’s attention         | 2018  |
| Datasets             | [7]       | Datasets for cognitive load inference using wearable sensors and psychological traits | 2020  |
| Datasets             | [24]      | Multilevel monitoring of activity and sleep in healthy people         | 2020  |
| Datasets             | [26]      | K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations | 2020  |
| Datasets             | [27]      | Toadstool: A dataset for training emotional intelligent machines playing super mario bros | 2020  |
| Machine Learning     | [16]      | Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study | 2019  |
| Machine Learning     | [40]      | Comparison of machine learning techniques for psychophysiological stress detection | 2016  |
| Machine Learning     | [8]       | Deep neural networks for human activity recognition with wearable sensors: Leave-one-subject-out cross-validation for model selection | 2020  |
| Machine Learning     | [38]      | Predicting stress in teens from wearable device data using machine learning methods | 2020  |
| Machine Learning     | [12]      | Comparison of regression and classification models for user-independent and personal stress detection | 2020  |
| Machine Learning     | [29]      | A review on mental stress detection using wearable sensors and machine learning techniques | 2021  |
| Machine Learning     | [7]       | Datasets for cognitive load inference using wearable sensors and psychological traits | 2020  |
| Machine Learning     | [15]      | A Sensitivity Analysis of Biophysiological Responses of Stress for Wearable Sensors in Connected Health | 2021  |
| Machine Learning     | [14]      | An Advanced Stress Detection Approach based on Processing Data from Wearable Wrist Devices | 2021  |
| Machine Learning     | [11]      | Monitoring stress with a wrist device using context                   | 2017  |
Table 2: Studies included in this review.

| Topic                  | Reference | Paper                                                                 | Date  |
|------------------------|-----------|-----------------------------------------------------------------------|-------|
| Machine Learning       | [20]      | Acute stress state classification based on electrodermal activity      | 2021  |
|                        |           | modeling                                                             |       |
| Machine Learning       | [17]      | Stress Detection from Multimodal Wearable Sensor Data                  | 2020  |
| Machine Learning       | [10]      | Objective Measurement of Physician Stress in the Emergency Department  | 2020  |
|                        |           | Using a Wearable Sensor                                               |       |
| Machine Learning       | [37]      | Evaluating the Reproducibility of Physiological Stress Detection       | 2020  |
|                        |           | Models                                                                |       |
| Machine Learning       | [21]      | Detection and Characterization of Physical Activity and Psychological  | 2020  |
|                        |           | Stress from Wristband Data                                            |       |
| Future Research        | [41]      | The quantified self: Fundamental disruption in big data science        | 2013  |
|                        |           | and biological discovery                                              |       |
| Future Research        | [42]      | Deriving a cortisol-related stress indicator from wearable skin        | 2020  |
|                        |           | conductance measurements: Quantitative model experimental validation  |       |
4. Results

Having reviewed the wide literature, we came across four main findings, which form the structure and discussions of our paper. Firstly, we found a general focus on the physiological aspects of stress and how specific biomarkers can potentially indicate periods of acute or chronic stress. Secondly, we found that the concept of The Quantified Self is gaining momentum \[41\] and popularity with the availability of affordable and non-intrusive measuring devices such as the Fitbit Sense [1], Empatica E4 [2] and Oura Ring 3 [43]. Thirdly, a distinct gap exists in the availability of public datasets where stress response was measured and labeled for periods of longer than twenty-four hours. Generally, experiments were conducted and data recorded for periods of less than 4 hours and a small (<30) number of subjects. Fourthly, we found that a wide variety of study-specific testing and measurement protocols were employed that will limit the re-use of the data for further experimentation and study, including building generic stress response predictor models using machine learning that generalize well on unseen data. Below, we discuss our results in more details.

4.1. Wearable Devices for Stress Measurement

Wearable devices such as smart watches and health monitors have become increasingly popular in the last few years. Advances in hardware such as component miniaturization have enabled more technological features to be embedded into ever shrinking devices at lower cost \[44\]. Today wearable devices are becoming as popular as smart phones, either as standalone health monitors or to complement existing consumer devices. The healthcare sector is likely set to see significant growth and investment in wearable technologies, and healthcare companies are investing significant amounts of time and financial resources into both hardware and software solutions built around wearable technology. 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 \[44\].

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. However, in this review, our focus was limited to devices that are capable of stand-alone monitoring as listed in Table 3, without the need for an additional harness or pairing with a secondary device, as this would limit the usefulness for study outside of a stricter laboratory setting.

We further limited this review to devices that are capable of recording multiple biomarkers known to be robust indicators of elevated levels of stress. Siirtola [45] 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. [46] concluded that EDA is a robust, reliable, not-subjective psycho-physiological biomarker of psychological stress within subjects, but not always between. Greco et al. [20] 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 5.1. Devices reporting stress based on only a single biomarker (typically HR or HRV) were therefore excluded.

Of the five devices included in this review, four are FDA-approved and wrist-worn, with the Oura Ring 3 is worn on the finger. All five devices provide biomarkers for at minimum HR, EDA and body temperature. The Empatica E4 would be considered medical-grade, with the remaining aimed at the consumer market. Two of the devices retail below US$500, while the Empatica E4 is priced at US$1,690 at the time of writing. The NOWATCH [47] and Study Watch were not available for sale at the time of review. Uniquely to the NOWATCH, Cortisol response is modeled [42] through the EDA signal to provide a cleaner and more robust estimation of stress. Battery life ranged from 30 hours for the Empatica E4 to 14 days for the NOWATCH, an important factor when performing long-running, continuous experiments.

4.2. Wearable Device Datasets for Stress Measurement

A number of datasets are publicly available containing sensor data recorded using a variety of devices including those in Table 3. Among these, Empatica provides an easy to use platform and interface for downloading recorded sessions which is of great benefit to researchers when choosing a wearable
device to utilize within a study, and as a wrist-worn, FDA-approved device we found it the most suitable, and widely-studied of the devices reviewed. Table 4 provides a summary of the datasets recorded using Empatica E4 alone or in combination with another measurement device. The SWELL dataset is the only dataset that used the Mobi device, but is included as it was utilized alongside the WESAD dataset by Nkurikiyeyezu et al. [18] during experimentation and contains both HR and EDA biomarkers similar to the Empatica E4. All these 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. AffectiveROAD and Toadstool 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 datasets reviewed were labeled 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.
|                  | Fitbit Sense | Empatica E4 | Study Watch | NOWATCH | Oura Ring 3 |
|------------------|--------------|-------------|-------------|---------|-------------|
| Battery Life     | 6 Days       | 30 Hours    | 1 Week      | 14 Days | 7 Days      |
| Placement        | Wrist        | Wrist       | Wrist       | Wrist   | Finger      |
| FDA Approved     | Y            | Y           | Y           | Y       | N           |
| Heart Rate       | Y            | Y           | Y           | Y       | Y           |
| Body Temperature | Y            | Y           | N           | Y       | Y           |
| Oxygen Saturation| Y            | Y           | Y           | Y       | Y           |
| EDA Sensor       | Y            | Y           | Y           | Y       | Y           |
| 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                     |
4.3. Machine Learning 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 5 lists some of the reviewed papers and the machine learning techniques utilized. Note that, while there are many more published papers on the general use of a machine learning techniques for stress detection using wearable devices, we judged the studies based on our observation and their contribution to the field to only include those that add new knowledge to the field.

Regardless of the utilized model or algorithm, a commonly used machine learning pipeline illustrated in Figure 3 is usually followed. In the beginning, data is pre-processed to transform into a format suitable for computerized processing (if required). At a later step, features may be created based on elements contained within the input data, in a process called feature engineering. A suitable machine learning algorithm is then selected (see Figure 1) depending on the problem at hand and desired output. This is followed by tuning algorithmic parameters to ensure maximum performance where available. The algorithm is then trained to learn the input data features and finally the predictive performance of the algorithm is analyzed. The algorithm selection, parameter optimization, and training steps may be repeated a number of times until performance is deemed to have reached the objective. It is worth noting that, some recent machine learning algorithms such as deep learning models, do not usually require a specific feature engineering step, as they are capable of automatically extracting useful features in the input data without domain expert knowledge.

In the following subsections, we provide a discussion on the different steps of machine learning pipeline for stress detection from wearable sensor data, and analyze how previous works have performed these steps, noting their strength and limitations.
Figure 3: A typical machine learning pipeline.
| Paper | Year | Model                  | Dataset | Accuracy | Subjects | Features | Cross Validation | Window          |
|-------|------|------------------------|---------|----------|----------|----------|------------------|-----------------|
| Comparison of Machine Learning Techniques for Psychophysiological Stress Detection | 2016 | Bayesian networks      | Custom  | 84.60%   | 20       | 22       | LOSO             | 30s, 29s overlap |
| Monitoring stress with a wrist device using context | 2017 | SVM                    | Custom  | 71.00%   | 5        | 6        | LOSO             | 6min            |
| Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection | 2018 | Random Forest, LDA, AdaBoost | WESAD  | 93.00%   | 15       | 82       | LOSO             | 0.25s, 5s, 60s   |
| Continuous Stress Detection Using Wearable Sensors in Real Life: Algorithmic Programming Contest Case Study | 2019 | Neural Network, Random Forest | Custom  | 88.20%   | 21       | 17       | 10-Fold          | 2min, 20min     |
| The Effect of Person-Specific Biometrics in Improving Generic Stress Predictive Models | 2019 | Random Forest, ExtraTrees | WESAD, SWELL | 93.90% | 15, 25   | 94       | 10-Fold          | 5min, 10min     |
| Datasets for Cognitive Load Inference Using Wearable Sensors and Psychological Traits | 2020 | XGBoost                | Snake, CogLoad | 82.00%  | 23       | 23       | LOSO             |                 |
| Comparison of Regression and Classification Models for User-Independent and Personal Stress Detection | 2020 | Bagged tree based ensemble | AffectiveROAD | 82.30% | 9        | 119      | LOSO             | 60s, 0.5s overlap |
| Predicting Stress in Teens from Wearable Device Data Using Machine Learning Methods | 2020 | Random Forest          | Custom  | 89.40%   | 8        | 756      | 10-Fold          |                 |
| Objective stress monitoring based on wearable sensors in everyday settings | 2020 | K-nearest-neighbor     | Custom  | 94.55%   | 17       | 25       | 10-Fold          | 60s             |
| Stress Detection from Multimodal Wearable Sensor Data | 2020 | RF                     | WESAD   | 92.00%   | 15       | 4        | 60/40 Split      | 0.25min         |
| Objective Measurement of Physician Stress in the Emergency Department Using a Wearable Sensor | 2020 | Naive Bay              | Custom  | 64.50%   | 8        | 30       | 10-Fold          | 20min           |
| Detection and Characterization of Physical Activity and Psychological Stress from Wristband Data | 2020 | LDA                    | Custom  | 98.30%   | 24       | 2216     | 10-Fold          |                 |
| A Sensitivity Analysis of Biophysiological Responses of Stress for Wearable Sensors in Connected Health | 2021 | Logistic Regression    | WESAD   | 85.71%   | 14       | 5        | 14-Fold          | 60s             |
| An Advanced Stress Detection Approach based on Processing Data from Wearable Wrist Devices | 2021 | Neural Network         | WESAD   | 85.00%   | 15       | 14       | 15-Fold          |                 |
| Acute stress state classification based on electrodal activity modeling | 2021 | SVM, RF                | Custom  | 94.62%   | 65       | 14       | LOSO             |                 |
4.3.1. 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 4Hz, while the HR signal is sampled at 1Hz. 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 1Hz 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, [7, 17, 18, 38] 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 stress wearable data is dataset balancing that can be done in different ways. For instance, Nkurikiyeyezu et al. [18] balanced the recorded sensor data by randomly discarding some samples from the majority (stressed) class, and further applied logarithmic, square root, and Yeo-Johnson transformations to ensure a Gaussian distribution, as required by their used linear regression model. Can et al. [16] also performed class-balancing through random down-sampling of the majority class (non-stressed observations) to match the minority class (stressed observations).

Another common pre-processing step is data standardization and normalization. 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, normalization and standardization were utilized by [7, 15, 18, 21], with [7] experimenting on both raw and standardized data, and finding that standardization offered improved predictive performance across all 10 machine learning algorithms tested.

Another usual preprocessing step on biomedical signals such as stress-related biomarkers collected by wearable devices is filtering. It is done to reduce out-
liers and any potential noise. For instance, [23] applied a 5Hz low-pass filter on the raw EDA signal, [15] applied a high-pass filter on the raw EDA signal, while [14, 18, 19] applied a 4Hz fourth-order Butterworth low-pass filter, followed by a moving average filter, to reduce outliers and remove noise from EDA sensor signals.

In examining the correlation between the EDA and HR biomarkers (Figure 4) and the stress label metric within the WESAD dataset (used in experiments by [14, 15, 17, 18, 23], there are noticeable correlation between both biomarkers and the stress metric. For the SWELL dataset, however, (Figure 5), also used by Nkurikiyeyezu et al. in [18], the correlation between EDA and the stress metric is virtually non-existent.

This observation highlights the challenges researchers face when selecting a suitable dataset for training a machine learning model that will then be applied on new, unseen data, collected from completely different experimental conditions. Normalization, standardization and outlier management is therefore crucial prior to implementing further feature-engineering [37]. Can et al. [16] utilized the accelerometer and temperature sensor biomarkers to develop a Support Vector Machine (SVM) capable of detecting noise artifacts within the EDA signal for subsequent removal, with an accuracy rate of 95%.
Figure 4: Correlation between biomarkers and stress label for WESAD dataset.
Figure 5: Correlation between biomarkers and stress label for SWELL dataset.
4.3.2. 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. [48] 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, while 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. [48] were frequently applied in the stress detection studies reviewed. Ten of the reviewed approaches [11-13, 15, 16, 18-21, 23] 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 [49], 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. [16] 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 [16] and [7] found that sliding windows ranging between 10 and 17.5 minutes produced better detection accuracy, with [16] further noting that different machine learning algorithms relied on different window sizes, an important factor to consider for future research.

Jin et al. [38] 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. In another study, Han et al. [19] engineered a total of 25 features from the ECG, PPG and galvanic skin response (GSR) biomarkers and then used greedy step-wise feature selection to identify the top 6 features considered most useful for their specific machine learning models. Step-wise feature selection were similarly utilized by Gjoreski et al. [11].
Notably, important engineered features included the proportion of the successive normal to normal beat intervals that differ more than 20 ms (ECG), ECG standard deviation, PPG on the Low frequency band (0.04–0.15 Hz) and PPG low frequency over high frequency band. Interestingly, these features were also engineered and utilized in experiments by [13], [23] and [16]. Smets et al. [13] utilized a correlation-based feature selection technique, and identified 4 key features for their general, subject-independent models, namely mean heart rate, the tonic and phasic components (from ECG and SCL) and the tonic and phasic components from the GSR biomarker. Gjoreski et al. [11] 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. [15] found features based on HR and respiratory rate to be the most important.

It is important to understand that engineered features are still representative of their origin feature, being sensor biomarkers recorded by wearable devices such as EDA, HR, TEMP and BVP. Current wearable devices are limited to these sensors, and this directly limits research into applying machine learning to predict stress to a finite set of biomarkers, and any derived features. New sensors could be developed in the future that expands on these biomarkers, opening up new potential avenues for stress prediction using machine learning.

4.3.3. Algorithm Selection

Of the 15 ML-based stress detection studies reviewed, we noted the use of 16 different machine learning 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 Neighbour (kNN), Multi-layer Perceptron (MPL), Multi-task 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. [18] experimented on a single method (Random Forest) while [45] 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, [10, 11, 13, 16, 19, 23] utilized 5 to 7 different algorithms and compared the stress prediction accuracy of each, with the highest performing models listed in Table 5.

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. [11] and Kaczor et al. [10].

Table 6 provides a summary of the various machine learning models tested in our reviewed studies. Interestingly, despite the use of deep-learning models in the wide healthcare domain [44], their use in stress detection and classification is limited. This can be due to the simple nature of the wearable signals, which can be analyzed using classical ML algorithms. Nonetheless, similar techniques to those used in processing other time-series biomedical signals such as EEG for seizure detection [50] can be explored. Among the reviewed works, only one study [14] reported using a Fully Convolutional Network (FCN) and the well-known Convolutional Neural Network to achieve a predictive accuracy rate of 85%, which is lower, when compared to classical ML methods.

It is worth noting that all the studies reviewed utilized supervised classification algorithms (Figure 1), apart from Siirtola et al. [45] that specifically tested the predictive accuracy of supervised classification algorithms compared to supervised regression algorithms, and found that supervised regression algorithms outperform their classification counterparts. Mishra et al. [37] similarly concluded that classification algorithms may not be the most suitable approach when applying machine learning to the problem of detecting stress.
Table 6: Machine learning models applied in reviewed literature.

| Paper                                                                 | LR | SVM | DT | RF | BN | PCA + LDA | PCA + SVM | kNN | MLP | Boosting | NB | Bagging | XGB | MTL | LDA | NN |
|----------------------------------------------------------------------|----|-----|----|----|----|-----------|-----------|-----|-----|----------|----|---------|-----|-----|-----|----|
| Comparison of Machine Learning Techniques for Psychophysiological Stress Detection [13] | •  | •   | •  | •  | •  | •         | •         | •   | •   | •         | •  | •       |     |     |     |    |
| Monitoring stress with a wrist device using context [11]            | •  | •   | •  | •  | •  |           |           | •   | •   | •         | •  | •       |     |     |     |    |
| Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection [23] | •  |     | •  | •  | •  |           |           | •   | •   | •         | •  | •       |     |     |     |    |
| The Effect of Person-Specific Biometrics in Improving Generic Stress Predictive Models [18] | •  |     | •  | •  | •  |           |           | •   | •   | •         | •  | •       |     |     |     |    |
| Continuous Stress Detection Using Wearable Sensors in Real Life: Algorithmic Programming Contest Case Study [16] | •  | •   | •  | •  | •  |           |           | •   | •   | •         | •  | •       |     |     |     |    |
| Stress Detection from Multimodal Wearable Sensor Data [17]          | •  | •   | •  | •  | •  |           |           | •   | •   | •         | •  | •       |     |     |     |    |
| Predicting Stress in Teens from Wearable Device Data Using Machine Learning Methods [38] | •  | •   | •  | •  | •  |           |           | •   | •   | •         | •  | •       |     |     |     |    |
| Objective stress monitoring based on wearable sensors in everyday settings [19] | •  | •   | •  | •  | •  |           |           | •   | •   | •         | •  | •       |     |     |     |    |
| Objective Measurement of Physician Stress in the Emergency Department Using a Wearable Sensor [10] | •  | •   | •  | •  | •  |           |           | •   | •   | •         | •  | •       |     |     |     |    |
| Detection and Characterization of Physical Activity and Psychological Stress from Wristband Data [21] | •  | •   | •  | •  | •  |           |           | •   | •   | •         | •  | •       |     |     |     |    |
| Datasets for Cognitive Load Inference Using Wearable Sensors and Psychological Traits [7] | •  | •   | •  | •  | •  |           |           | •   | •   | •         | •  | •       |     |     |     |    |
| Comparison of Regression and Classification Models for User-Independent and Personal Stress Detection [12] | •  | •   | •  | •  | •  |           |           | •   | •   | •         | •  | •       |     |     |     |    |
| An Advanced Stress Detection Approach based on Processing Data from Wearable Wrist Devices [14] | •  | •   | •  | •  | •  |           |           | •   | •   | •         | •  | •       |     |     |     |    |
| Acute stress state classification based on electrodermal activity modeling [20] | •  | •   | •  | •  | •  |           |           | •   | •   | •         | •  | •       |     |     |     |    |
| A Sensitivity Analysis of Biophysiological Responses of Stress for Wearable Sensors in Connected Health [15] | •  |     | •  | •  | •  |           |           | •   | •   | •         | •  | •       |     |     |     |    |
4.3.4. 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 hyperparameter 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 have been developed to scan the full parameter search space in a reasonable amount of time to determine the optimal combination. Some of these methods include:

- Grid search - an exhaustive search through a set of manually specified hyperparameter values.
- Random search - randomly searching the grid space instead of performing an exhaustive search.
- Bayesian optimization - builds a probabilistic model to map parameters to an objective function.
- Gradient-based optimization - specifically for neural networks, computes the gradients with respect to hyperparameters and optimizes them using the gradient descent algorithm.
- Evolutionary optimization - uses genetic or evolutionary algorithms to search the hyperparameter space.

Of the stress-related studies reviewed, we noted [13] restricted the hyperparameter of the estimators count used in their Random Forest model to 20, while [38] performed a grid search with estimators set at 500. In [18], 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 [23], 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. [19] 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. [21] utilized Bayesian optimization techniques for feature-selection.

Unlike the aforementioned works, Gjoreski et al. [7] 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. [16], 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.

### 4.3.5. Model Training and Validation

As previously shown in Table 6, several of the previous works [10, 11, 13, 16, 19, 23] trained and evaluated 5 different machine learning models and compared their performance, while [7] utilized 8 different algorithms. In addition, [18] compared predictive performance using a single Random Forest model in order to establish a baseline for testing the effect of person-specific bio-metrics in improving generic stress prediction models, using both the WESAD and SWELL datasets.

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 [10, 14, 17, 20, 21, 23, 27, 39]; (ii) self-reporting via questionnaires [7, 11, 15, 24, 26]; and (iii) labeling by a third-party observer, who observes subjects’ response to a situation and numerically scores/grades the level of stress observed [26, 25].

As highlighted in Table 5, the best performing models from each experiment achieved at least 64.5% accuracy, with [17, 21, 23] reporting binary classifica-
tion 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 measure 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. [45] attempted to model how high this relationship is (using a regression algorithm instead of classification), while Umematsu et al. [51] 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. K-fold cross-validation has a single parameter called $K$ that refers to the number of groups that a given data sample is to be split into. Leave-one-subject-out (LOSO) cross validation is K-fold cross validation taken to its logical extreme, with $K$ equal to $N$, the number of data points in the dataset. That means that $N$ separate times, the machine learning algorithm is trained on all the data except for one point, and a prediction is made for that point.

Among the studies reviewed, [7, 11–14, 20, 23] utilized LOSO cross-validation, while [10, 15, 16, 18, 19, 21, 38] utilized K-fold cross-validation with $K=10$. In addition, [7] utilized both LOSO and K-fold cross-validation, with $K=5$. All studies reviewed approached stress prediction as a binary classification problem apart from [45], 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.

4.3.6. Performance Analysis
Selecting the most appropriate performance metric is crucial for evaluating machine learning models. A wide variety of metrics are available depending on the problem, for example classification or regression type problems. The experiments reviewed utilized and reported a number of different evaluation
metrics including F-score [7, 10, 23], classification accuracy [10, 11, 13, 17, 19, 21], Area Under the Curve (AUC) [38], Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) [18]. For comparisons among the reviewed studies, here we only investigate their achieved classification accuracy.

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:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

To determine classification accuracy, Can et al. [16] and Kaczor et al. [10] divided the training data into 3 classes: stressed, stressed with provided experimental content, and non-stressed, reporting a classification accuracy rate of 90.4%. Gjoreski et al. [7] 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 classification accuracy rate of 94.55%, while Nkurikiyeyezu et al. [18] and Schmidt et al. [23] reported binary classification accuracy rates of 93.9% and 93% respectively. Jin et al. [38] 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 accuracy rates can be high even if the predictions for a minority class is 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.

The experiments found to use machine learning for stress detection were performed between 2016 and 2021, and Figure 6 details the reported accuracy metric achieved for these experiments. Interestingly, similar rates of accuracy were achieved regardless of validation technique utilized. The highest reported accuracy rate of 98.30% was achieved by Sevil et al. [21] utilizing LDA with 10-Fold cross-validation on a dataset containing sensor biomarkers of 24 subjects. The dataset used however is not publicly available, complicating the ability to reproduce the results.

Figure 6 shows an overall upward trend (see the orange bars) in the highest accuracy achieved from the first machine learning work for stress detection from wearable data in 2016, to the recent works in 2021. Most of the reported
accuracy rates are concentrated between 85% to 95%, and we note a general lack of using deep-learning or ensemble techniques to combine the predictive strengths of different algorithms, into a single predictor model.

With the exception of studies done by [42], and the results being incorporated into the NOWATCH [47] wearable device, there appears to be a lack of studies exploring the interaction between stress biomarkers recorded using wearable sensors and other known physiological stress biomarkers such as cortisol.

![Figure 6: Predictive accuracy over time. Here, orange bars show the highest accuracy achieved in each year, with the other colors showing the subsequent accuracy values reached in each year.](image)

In virtually all the reviewed studies the number of subjects were less than 30. None of these studies apart from Mishra et al. [37], tested generalization of the resulting model on a totally unseen, new dataset. The results from LOSO or K-Fold cross-validation techniques were instead used to validate
their model generalization capability.

Figure 7 compares the accuracy rates reported for each cross-validation technique applied. LOSO cross-validation consistently reported lower rates of accuracy compared to K-fold, however since these experiments were not all performed on the same dataset, using the same machine learning model with exact same hyperparameter selection, no definitive conclusions can be reached to support either cross-validation technique.

Interestingly, as shown in Figure 8, there appears to be no obvious correlation between the number of subjects included in the study with the reported accuracy rate. However, as noted previously, the number of subjects included in all studies reviewed are relatively low, at 25 or less, preventing any definitive conclusions to be reached.
5. Discussion

In order to build a reliable stress detection system, it is important to understand that stress is primarily a physiological response to a stimulus triggered by the sympathetic nervous system (SNS). During this response a mixture of hormones like cortisol or adrenaline are released, leading to increased breathing/heart rate and muscle tension. These physiological changes prepare the organism for a physical reaction (“fight-or-flight”) \[23\]. It is further important to note that these changes take place over time, more so for chronic stress, leading to the understanding that wearable sensor data is in effect, a time series of observations of biomarkers that could potentially indicate occurrences of elevated levels of stress in an individual.

In order to build a robust machine learning model capable of accurately detecting stress: (i) sensor biomarker data needs to be valid and sufficiently varied to capture a wide spectrum of potential physiological stress response;
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) A sufficient volume of data is required to ensure a high level of statistical power is available, thereby ensuring (iv) model generalization occurs in order to accurately predict on new, unseen data. The discussion of this review is therefore focused on those four key requirements.

5.1. Validity of sensor biomarkers

The devices included in this study offer a varying array of sensors potentially useful in detecting elevated levels of stress via HR, HRV and EDA signals, measured across a time interval, and of these devices the Empatica E4 has been studied the most in-depth. In addition, sweat sensing is at the forefront of wearable stress detection currently in development [28] 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, [52] 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, but that physiological changes during a workday can be tracked by the Empatica E4 wearable against more major, sustained stressors [52]. In [53], 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. [54] and Sevil et al. [21] that found the Empatica E4 can be severely compromised by motion artifact, resulting in a high percentage of missing data across all conditions except seated and supine baselines, and calls into question the E4’s efficacy as an HRV measurement tool in most in-vivo conditions. In contrast, Greco et al. [20] found EDA to be a good marker of stress when features are engineered based on its phasic and tonic components. In [12] the authors predominantly focused on comparing regression vs. classification models using the AffectiveROAD dataset [25]. This dataset contains sensor recordings for both left and right hands.
of the test subjects. To ensure consistent, comparable results, [12] 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 ([2]) that newer studies have shown substantial differences in the EDA signal between the dominant and non-dominant hand.

5.2. Labeling protocol
In terms of labeling protocol and methodology, [12] questioned the accuracy of self-reporting of perceived levels of stress experienced, which was previously questioned by [55], who noted that study subjects are less likely to report on states less socially desired. Accurate labeling of stress/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 [55].

5.3. Lack of statistical power
It is common to design behavioral science experiments with a statistical power of 80% or higher [56], which reduces the probability of encountering a Type II error of up to 20% [57]. Statistical power has three parts: effect size (a statistical measure), sample size (number of observations or participants) and significance (typically 0.05 [56]). 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.

One of the recurrent questions psychology researchers ask is: "What is the minimum number of participants I must test?" [58]. 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 [58]. However, when applying a two-tailed power test with a correlation coefficient of 0.5 [56] and an assumed significance level of $\alpha=0.05$, we found that at least 34 test subjects would be required to achieve 80% power.

Considering the small number of subjects contained in the datasets utilized in the experiments reviewed in this paper, 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. [59] 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 samples size contains sufficient power.

Overall, the machine learning models and experiments reviewed do not specifically detail whether power analysis were performed, or what is the statistical power obtained given the number of subjects included within their individual datasets. Considering the small number of subjects within most of these datasets, this can be of a concern, and the accuracy scores reported are unlikely to be achieved on new, unseen data.

5.4. Lack of generalization
Of the machine learning approaches reviewed in Table 5, all were built as classification models, either binary or multi-class, for classifying detected
levels of stress (stressed vs. not-stressed) while [18] and [12] opted for regression models, with [12] specifically comparing the performance of classification against regression models, and concluding that regression models vastly outperform classification models in terms of predictive accuracy. Importantly, [12] noted that the real benefit of regression over classification models is that they can be used to estimate the level of stress detected.

In [13, 16, 18, 38] and [45] the authors compared the performance of general models to personalized models. General models utilized the same biomarkers as model features across all subjects, while personalized models selected the biomarker features with the highest levels of correlation to elevated levels of stress, per individual subject. Apart from [45], all studies concluded that personalized models delivered higher levels of accuracy, with [13] noting that personalized models are able to give more insight into the personal physiological stress response indication of each subject’s principal stress physiology.

Reported accuracy metrics ranged from 64.5% (lowest) to 98.3% (highest). Of the 15 experiments reviewed, [18] and [7] utilized 2 independent datasets while the others used a single dataset and experimental setup. Considering the small number of subjects used in the reviewed experiments, it is doubtful that any of their proposed models, trained on their customized datasets, can generalize well on a totally new, unseen dataset. An approximate method of measuring generalization is to use cross-validation. We note that in the reviewed literature, usually two key cross-validation schemes were employed during model training, i.e. 10-Fold and Leave-One-Subject-Out, however this is simply not the only parameter involved when training machine learning models as noted by [23]. Generalization will likely only be achieved with a large number of test subjects and the proper identification of key sensor biomarkers that correlate well with elevated levels of stress, combined with suitable feature-engineering techniques and a robust study protocol.

Towards this end, Mishra et al. [37] examined the generalization of machine-learning models for stress detection and analyzed data from 90 participants across 4 independent controlled studies, using two different types of sensors, with different study protocols and research goals. Based on their findings, the following conclusions were reached:
• It is not yet feasible to deploy real-time, in-the-moment stress detection that works immediately out-of-the-box. This is partly due to models requiring normalization of data prior to training to remove any subject-specific traits, and new, unseen data would not be included during the normalization step, resulting in poor predictive power.

• Applying a model on new data recorded outside of a controlled, laboratory environment would be problematic, as factors that could affect the sensor biomarkers such as stress context, caffeine intake or tobacco use would be outside of experimental controls.

• Stress is not a binary problem, also noted by Siirtola et al. [12], and a regression approach is likely to yield a better measure of the interaction of sensor biomarkers, requiring further interpretation and assessment by an expert.

5.5. 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.

• Machine learning is being utilized heavily to model, measure, predict and understand the data recorded using wearable devices and their correlation to elevated levels of stress.

• Existing work have predominantly used small datasets acquired in a single experimental setup with varying labeling protocols, bringing into question the generalization ability of their built machine learning models.
5.6. 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 EDA, HR and HRV data, including appraisal and scoring methodology. In [35], the authors concluded that, the appraisal process critically shapes an individual’s response to acute stress, while [60] detected lower EDA activity to 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, i.e. on the finger [43].

- Another significant challenge is the lack of 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.

The main identified lacunas in which future research work should concentrate are as follow:

- Development of a sufficiently large and varied datasets that can be utilized to design and build machine learning models capable of generalizing well on unseen data. This will likely occur as newer devices reach the market with open platforms for easily accessing the recorded sessions and data. Merging of existing datasets might also be possible where recorded biomarkers and experimental setup is similar, resulting in machine learning models that generalize better.
• Development of a robust stress detection system to identify elevated levels of stress in non-laboratory conditions, while being easy to implement and utilize for most researchers. Robust models depend on robust data, and accuracy and ability to generalize will further improve as more data becomes available to machine learning researchers. For instance, more vendors need to allow end users to download their raw sensor biomarker data, similar to the platform provided by Empatica.

• Identifying methods for predicting acute and episodic versus chronic stress over longer periods of wearable device use. The machine learning models reviewed predominantly focused on identifying acute stress events, while long-term, chronic stress has the highest impact on overall health. More research needs to be done that focus on detecting and measuring chronic stress, dependent on the availability of larger datasets spanning longer time-frames.

6. Conclusion
The main objective in automated stress detection and measurement is to develop a highly accurate model that is effective and can generalizing well on new, unseen data. The review presented here synthesized the literature and presented important information about the previous studies concerned with utilizing wearable devices to detect and measure elevated levels of stress. In particular, we reviewed and analyzed the datasets built and utilized, the machine learning techniques applied, their advantages, limitations, claimed accuracy and known issues. 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.

References
[1] Fitbit.com, Fitbit (1 2022).
URL http://www.fitbit.com

[2] Empatica, Empatica — medical devices, ai and algorithms for remote patient monitoring (1 2022).
URL http://www.empatica.com
[3] 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 Research 153 (2019) 79–82. doi: 10.1016/j.eplepsyres.2019.02.007.

[4] J. Tang, R. E. Atrache, S. Yu, U. Asif, M. Jackson, S. Roy, M. Mir-momeni, 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. doi: 10.1111/epi.16967.

[5] F. Onorati, G. Regalia, C. Caborni, W. C. LaFrance, A. S. Blum, J. Bidwell, P. D. Liso, R. E. Atrache, T. Loddenkemper, F. Mohammadpour-Tousserkan, 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, Frontiers in Neurology 12 (aug 2021). doi: 10.3389/fneur.2021.724904.

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

[7] 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, Applied Sciences 10 (11) (2020) 3843. doi: 10.3390/app10113843.

[8] 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. doi: 10.1109/access.2020.3010715.

[9] 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. doi: 10.3390/s21010052.

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

[11] M. Gjoreski, M. Luštrek, M. Gams, H. Gjoreski, Monitoring stress with
a wrist device using context, Journal of Biomedical Informatics 73 (2017)
159–170. doi:10.1016/j.jbi.2017.08.006.

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

[13] 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. doi:10.1007/
978-3-319-32270-4-2.

[14] M. Alshamrani, An advanced stress detection approach based on pro-
cessing data from wearable wrist devices, International Journal of Ad-
vanced Computer Science and Applications 12 (7) (2021). doi:10.
14569/ijacsa.2021.0120745.

[15] 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. doi:10.1109/access.2021.3082423.

[16] Y. S. Can, N. Chalabianloo, D. Ekiz, C. Ersoy, Continuous stress detec-
tion using wearable sensors in real life: Algorithmic programming con-
test case study, Sensors 19 (8) (2019) 1849. doi:10.3390/s19081849.

[17] F. I. Indikawati, S. Winiarti, Stress detection from multimodal wearable
sensor data, IOP Conference Series: Materials Science and Engineering
771 (1) (2020) 012028. doi:10.1088/1757-899x/771/1/012028.

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

[19] H. J. Han, S. Labbaf, J. L. Borelli, N. Dutt, A. M. Rahmani, Object-
ive stress monitoring based on wearable sensors in everyday settings,
Journal of Medical Engineering & Technology 44 (4) (2020) 177–189.
doi:10.1080/03091902.2020.1759707

[20] 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 Transactions on Affective Computing (2021) 1–10. doi:10.1109/taffc.2021.3055294.

[21] 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. doi:10.3390/signals1020011

[22] 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. doi:10.3390/diagnostics11030556

[23] 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. doi:10.1145/3242969.3242985

[24] 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 (2020). doi:10.13026/CERQ-FC86

[25] N. E. Haouij, J.-M. Poggi, S. Sevestre-Ghalila, R. Ghozi, M. Jaidane, AffectiveROAD system and database to assess driver’s attention, in: Proceedings of the 33rd Annual ACM Symposium on Applied Computing, ACM, 2018. doi:10.1145/3167132.3167395

[26] 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 (2020). doi:10.5281/ZENODO.3931963

[27] 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 (feb 2020). doi:10.31219/osf.io/4v9mp

46
[28] C. Samson, A. Koh, Stress monitoring and recent advancements in wearable biosensors, Frontiers in Bioengineering and Biotechnology 8 (sep 2020). doi:10.3389/fbioe.2020.01037

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

[30] J. Waxenbaum, V. Reddy, M. Varacallo, Anatomy, Autonomic nervous system, StatPearls [Internet]. Treasure Island (1 2021).

[31] D. Broom, Cortisol: often not the best indicator of stress and poor welfare, Physiology News (Summer 2017) (2017) 30–32. doi:10.36866/ pn.107.30

[32] J. S. Christensen, H. Wild, E. S. Kenzie, W. Wakeland, D. Budding, C. Lillas, Diverse autonomic nervous system stress response patterns in childhood sensory modulation, Frontiers in Integrative Neuroscience 14 (feb 2020). doi:10.3389/fnint.2020.00006

[33] S. Cohen, T. Kamarck, R. Mermelstein, A global measure of perceived stress, J Health Soc Behav 24 (4) (1983) 385–396.

[34] M. Chesnut, S. Harati, P. Paredes, Y. Khan, A. Foudeh, J. Kim, Z. Bao, L. M. Williams, Stress markers for mental states and biotypes of depression and anxiety: A scoping review and preliminary illustrative analysis, Chronic Stress 5 (2021) 247054702110003. doi:10.1177/24705470211000338

[35] 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, Frontiers in Neuroendocrinology 49 (2018) 146–169. doi:10.1016/j.yfrne.2018.03.001

[36] A. Danese, B. S. McEwen, Adverse childhood experiences, allostatic, allostatic load, and age-related disease, Physiology & Behavior 106 (1) (2012) 29–39. doi:10.1016/j.physbeh.2011.08.019

[37] V. Mishra, S. Sen, G. Chen, T. Hao, J. Rogers, C.-H. Chen, D. Kotz, Evaluating the reproducibility of physiological stress detection models,
Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 4 (4) (2020) 1–29. doi:10.1145/3432220

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

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

[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 Digital Medicine 1 (1) (dec 2018). doi:10.1038/s41746-018-0074-9

[41] M. Swan, The quantified self: Fundamental disruption in big data science and biological discovery, Big Data 1 (2) (2013) 85–99. doi:10.1089/big.2012.0002

[42] 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, Frontiers in Computer Science 2 (sep 2020). doi:10.3389/fcomp.2020.00039

[43] Oura ring (1 2022). URL http://www.ouraring.com

[44] 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. doi:10.1109/tbcas.2020.3036081

[45] P. Siirtola, Continuous stress detection using the sensors of commercial smartwatch, 2019, pp. 1198–1201. doi:10.1145/3341162.3344831

[46] 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, Fron-
tiers in Human Neuroscience 6 (2013). doi:10.3389/fnhum.2012.00349.

[47] Nowatch.com, Nowatch (1 2022). URL http://www.nowatch.com

[48] C. Guo, M. Lu, J. Chen, An evaluation of time series summary statistics as features for clinical prediction tasks, BMC Medical Informatics and Decision Making 20 (1) (mar 2020). doi:10.1186/s12911-020-1063-x.

[49] S. Kreibig, Autonomic nervous system activity in emotion: A review, Biological psychology 84 (2010) 394–421. doi:10.1016/j.biopsycho.2010.03.010.

[50] C. Lammie, W. Xiang, M. R. Azghadi, Towards memristive deep learning systems for real-time mobile epileptic seizure prediction, in: 2021 IEEE International Symposium on Circuits and Systems (ISCAS), 2021, pp. 1–5. doi:10.1109/ISCAS51556.2021.9401080.

[51] 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. doi:10.1109/EMBC44109.2020.9176706.

[52] 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, Behavior Research Methods 52 (2) (2019) 607–629. doi:10.3758/s13428-019-01263-9.

[53] 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, Frontiers in Behavioral Neuroscience 14 (aug 2020). doi:10.3389/fnbeh.2020.00148.

[54] 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 (may 2019). doi:10.31234/osf.io/t3gdz.
[55] P. Schmidt, R. Dürrchen, A. Reiss, K. V. Laerhoven, T. Plötz, Multitarget affect detection in the wild, in: Proceedings of the 23rd International Symposium on Wearable Computers, ACM, 2019. doi: 10.1145/3341163.3347741.

[56] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum Associates, 1988.

[57] J. Brownlee, A gentle introduction to statistical power and power analysis in python (Apr 2020).

[58] M. Brysbaert, How many participants do we have to include in properly powered experiments? a tutorial of power analysis with reference tables 2 (1) (2019). doi:10.5334/joc.72.

[59] J. J. Li, X. Tong, Statistical hypothesis testing versus machine learning binary classification: Distinctions and guidelines, Patterns 1 (7) (2020) 100115. doi:10.1016/j.patter.2020.100115.

[60] 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, Journal of Autism and Developmental Disorders 45 (2) (2013) 576–588. doi:10.1007/s10803-013-1996-3.