# Emotion Recognition for Everyday Life Using Physiological Signals From Wearables: A Systematic Literature Review

Stanisław Saganowski, Bartosz Perz, Adam G. Polak, *Member, IEEE* and Przemysław Kazienko, *Senior Member, IEEE*

**Abstract**—Smart wearables, equipped with sensors monitoring physiological parameters, are becoming an integral part of our life. In this work, we investigate the possibility of utilizing such wearables to recognize emotions in the wild. In most reviewed papers, the authors apply a similar procedure consisting of participant recruitment, stimuli preparation and annotation, signal collection and processing, self-assessment, and machine learning model learning and validation. Besides, we identified seven emotion recognition scenarios and analyzed the transition from psychological models to machine learning tasks.
Even though the majority of the research was performed in the laboratory environment, we conclude that studies in the field are feasible. They require especially: (1) new self-assessment and triggering procedures adjusted to a real-life scenario, (2) more attention to the machine learning process, including suitable deep learning architectures, revision of the data imbalance problem, and subject-specific data processing, (3) adequate validation procedures, (4) consideration of the model generalizability vs. personalizability, (5) comfortable devices able to provide reliable measurements in motion. Additionally, more large-scale studies are necessary to increase result credibility. We also postulate actions toward replicability and comparability of the research.

**Index Terms**—emotion recognition, affective computing, systematic literature review, survey, review, field studies, validation, wearable, smartwatch, smart band, personal device

## 1 Introduction

Emotions drive most of our decisions [1], not only intuitive ones [2]. Therefore, they directly affect our everyday life. Most studies on emotion recognition conducted so far focused on participant (subject) reactions evoked by the prepared stimuli in the controlled environment (laboratory setup). In consequence, emotion identification in the real-life environment remains a significant challenge. Automatic emotion recognition using physiological signals from wearables has the potential to facilitate a breakthrough in healthcare, human-computer interaction, automotive, gaming, or e-learning. A system identifying depression would be a game-changer for mental-related problems and for the quality of life in general. People who have Autism Spectrum Disorder could gain a personal assistant to help them perceive and express emotions. Existing systems may be enhanced to respect emotions, providing better recommendations and user interfaces, more relevant content, better adjusted game difficulty, or emotionally aware cars.

In this paper, we focus on physiological signals that can be monitored using embedded sensors from popular wearables like smartwatches or wrist bands. Due to their unobtrusiveness and convenience, they facilitate emotion recognition in the real-life environment, a.k.a. *field studies* or *in the wild studies*.

It is difficult to provide a commonly agreed definition of emotion. We should rather consider a set of features that distinguish emotion from non-emotion [3]. Overall, *affect* is seen as a neurophysiological state that is consciously accessible but not directed at any specific entity. *Mood* is a lasting and not very intense sensation. Finally, a short, intense and directed feeling is described as *emotion* [4]. In this review, we refer to emotions interchangeably with affect. It is also clear that two affective computing topics: stress and emotion recognition, have become rather separate research lines [5]. Most researchers assume that emotions can be recognized and evaluated using objective behavioral or physiological signals confronted against subjective self-perception or emotional labels assigned to stimuli.

Recently, several survey studies on emotion recognition have been published. They are related to different modalities, e.g. facial expression [6], speech [7], electroencephalogram (EEG) [8], multimodal approaches [9], [10], and also physiology-oriented ones [4], [11]–[14]. Meyer et al. [10] conducted a systematic literature review to find mobile emotion measurement and recognition solutions. They distinguished four sources of emotional traces: face, speech, biofeedback, and gestures. Their numerical analysis of the literature confirms that physiology is the most *hot topic* in emotion recognition. Dzedzickis et al. [12] discussed types of sensors used to assess emotions: tracking brain and muscle activity, monitoring cardiac functioning, and skin parameters (conductance and temperature). Maria et al. [13] compared emotional modalities (speech, facial expression,

- *S. Saganowski, B. Perz, and P. Kazienko are with the Department of Artificial Intelligence, Faculty of Information and Communication Technology, Wrocław University of Science and Technology, Wrocław, Poland. E-mail: {stanislaw.saganowski, bartosz.perz, kazienko}@pwr.edu.pl*
- *A. G. Polak is with the Department of Electronic and Photonic Metrology, Faculty of Electronics, Photonics and Microsystems, Wrocław University of Science and Technology, Wrocław, Poland. E-mail: adam.polak@pwr.edu.pl*

*Manuscript received March XX, 2021; revised March XX, 2021.*

physiological signals) in terms of usefulness, limitations, and application area. They stated the multimodal approach would be the best for physiological signals. Shu et al. [14] covered all aspects of emotion identification: emotion models, stimuli, feature extraction, model training. They did not consider whether the research was conducted in the lab or field setup. Schmidt et al. [4] respected the surveyed studies' environment and concluded that emotion recognition outside the lab is much more difficult. They provided practical guidelines for designing and applying ecological momentary assessment in field studies.

In this survey, we go a step further and consider studies that are (or could be) placed in the field, i.e., they use devices and methods that allow us to recognize emotions in everyday situations. Following the systematic literature review (SLR) procedure, we cover all relevant research published up to date. For that purpose, we reviewed 3,051 papers and found only 34 (1.1%) meeting all of our research criteria.

In 2020, we published the work-in-progress of our SLR [15]. Since then, we have focused more on the applicability of emotion recognition solutions in everyday life scenario. We revised our initial inclusion/exclusion criteria and removed some papers, e.g., studies with EEG devices, as their applicability is limited to a stationary position.

This paper systematizes and summarizes the extensive topic of emotion recognition for everyday life. It is intended to serve as a source of knowledge for researchers willing to conduct emotion recognition (field) studies. It covers all research stages: emotional models, participants recruitment and analysis, stimuli, signal processing, self-assessment, machine learning model development and validation. It also provides a thorough analysis and discussion of crucial research components, as well as the most prominent research directions in the domain. Due to page limit, Supplementary Materials (Supp. Mat.) attached to the main article cover: SLR results, analysis of individual SLR papers, other research components, datasets and software tools, application examples, as well as all aspects related to lab studies.

The contribution of this paper is as follows:

1) **SLR**. Systematic literature review on emotion recognition from physiological signals suitable for field studies.
2) **Lab vs. field and their scenarios**. Comparison between lab and field studies along with identification of seven research scenarios.
3) **Emotional models $\Rightarrow$ ML models**. Analysis of transition from psychological models of emotions to machine learning (ML) problems.
4) **Ground truth acquisition**. Analysis of labeling in lab studies. Methods used to trigger self-assessment in the wild. Questionnaires used for self-assessment.
5) **Biosignals and features**. Analysis of biosignals and feature extraction methods.
6) **ML steps, validation and parameters**. Identification of crucial machine learning steps and methods. A unique analysis of validation methods and study parameters.
7) **Discussion**. A comprehensive discourse on the current approaches, challenges, and the most promising directions in the emotion recognition task.
8) **Emognition portal**. An open web platform with SLR results, resources, models, datasets, libraries, and other relevant knowledge [16].

## 2 Affective Computing and Emotion Recognition

The overall goal of affective computing is to make computers recognize, understand, express, and reproduce human emotions [17]. Its crucial component is affect recognition, which can be seen as a dynamic pattern recognition problem commonly solved by means of supervised machine learning. Usually, affect covers a wide range of psychological states especially emotions but also stress [18]–[22], anxiety [21], [23], [24], flow [25] or mood [20], [26]. In this survey, we will primarily focus on emotion recognition solutions; however, some other affects were also investigated in the papers we identified. Moreover, consideration of multiple affects resulted in multi-task deep learning methods that simultaneously recognize diverse affective states [21].

The principal motivation for emotion recognition from physiological signals are natural, biological signs of emotions [27], which have been investigated for dozens of years. Schlosberg claimed in 1954 that electrical skin conductance is a good measure of the extent of emotional arousal [28]. Ekman, Levenson, and Friesen showed in the 80s and 90s that reaction of the autonomic nervous system (ANS) to voluntarily produced emotion may be observed by means of physiological signals [29], [30]. In particular, they found some associations between six primary emotions and heart rate, finger temperature, and skin conductance. They believed it comes from the ANS functional specificity.

To develop effective methods for emotion recognition, we need to accumulate appropriate learning samples, later used to train reasoning models. For that purpose, emotions can be identified and evaluated from different points of view, which complement each other, Fig. 1:

1) Behavioral signals like facial expressions, speech or specific body movements;
2) Reaction of participant's organism (physiological signals), which is objective but may be contaminated by the individual's body condition and functioning (impacted by drugs or illnesses);
3) Subjective perception of the subject combined with their ability to define what they emotionally experience; commonly collected through subject's self-assessment;
4) External evaluation made by the domain experts while observing the subjects, e.g., an adult recognizing the state of the child [31];
5) Dedicated stimuli or activities that are expected to invoke certain emotions.

Many studies and solutions relay on the first perspective – behavioral patterns, i.e., emotion recognition from facial expressions [6] and speech [7] but also from the whole body [32] or eye movements [33]. The last three perspectives are used to provide emotion labels, which are assigned to the evaluated objective signals. They are subjective and consciously delivered by humans, i.e., (3) by their self-awareness of recent emotional experiences; (4) by evaluation of visible or audible signs interpreted by external peers (experts); and (5) by researchers themselves, who use stimuli or activities with the predetermined emotional labels.

In this paper, however, we focus on the second perspective – physiological signals that can be collected by means of pervasive sensors built into wearable devices

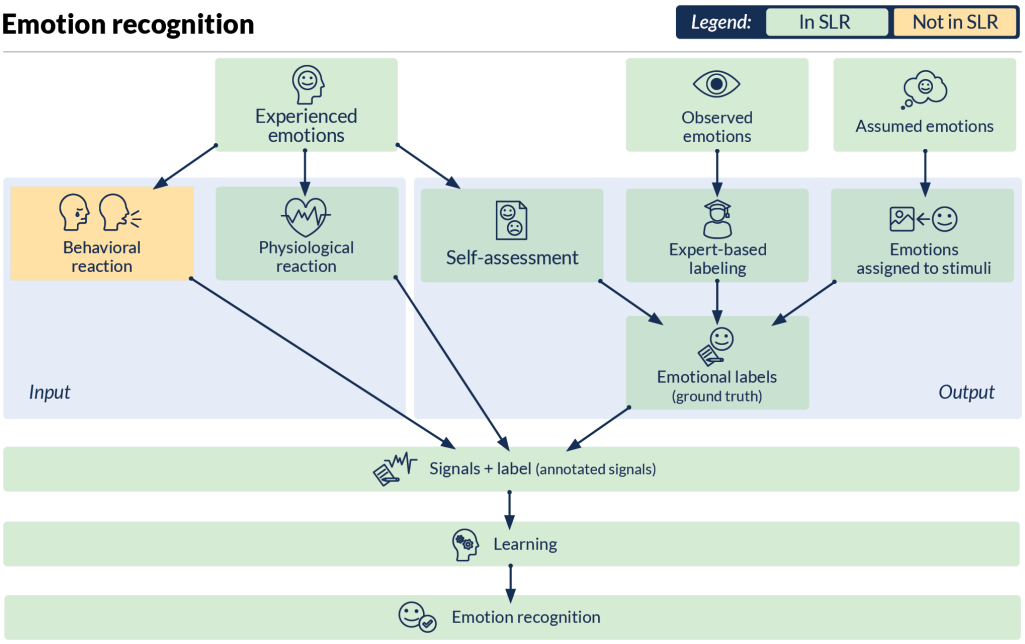


Fig. 1. General approaches to emotion recognition. Behavior-based methods like facial expressions or speech are not considered in this SLR.

TABLE 1
Usability of the most popular emotion recognition methods for field studies

| Feature | Facial expression | Speech | Physiology |
|---|---|---|---|
| Device | camera | microphone | wearable |
| Multipurpose devices | + | + | −/+ |
| Continuous monitoring in everyday life | − | − | + |
| Tracking during physical exercises | − | − | −/+ |
| Monitoring devices invisible | −/+ | + | −/+ |
| Main drawback | well visible face required | voice emission required | sensors touching body |

like smartwatches, wrist bands, smart rings, chest straps, or headbands. Since they are convenient, non-intrusive, wireless, and often multipurpose, they facilitate monitoring and recognition of emotions in everyday life (field studies).

Three main groups of methods for emotion recognition are briefly compared in Tab. 1 and discussed in Supp. Mat. Sec. 1. We believe that approaches exploiting physiological signals, especially those collected with the off-the-shelf wearables, are most suitable for field studies.

## 3 Systematic Literature Review

We have performed the *Systematic Literature Review* [34] (SLR) to answer the vital question: *Can wearables be used to recognize emotions in everyday life?*

To find relevant articles, we examined three databases: Scopus, Web of Science, and Google Scholar (via Publish or Perish). We asked the following query: *[emotion* or affective] and [wearable* or (smart watch) or iot or (personal device*) or (ambient intelligence) or (smart device*) or (smart band*)]*.

To exclude resources that are only vaguely related to our question, we evaluated each article with the set of inclusion and exclusion criteria. The inclusion criteria were as follows:

1) Personal devices/wearables were used to recognize (classify) various emotions. Device/wearable should enable emotion recognition in everyday life;
2) Personal device/wearable was described, or the description was available elsewhere;
3) At least one physiological signal was monitored and utilized to emotion recognition;

while the exclusion criteria were defined as:

1) The study was performed on a population less than five subjects;
2) Only a single emotion or its levels was considered;
3) None of the exploited devices was personal/wearable/portable;
4) The device had modules interconnected with cables, e.g., BioPac system where sensors were wired to the development board.

The potential solutions/systems/devices should be applicable to any possible case that might occur in daily life rather than to one or few specific scenarios. They should be free of constraints limiting the applicability to a particular condition, e.g., emotion recognition with a webcam, which works only if the subject is in front of the laptop.

Moreover, we did not include studies that focused on affective states lasting longer than a few minutes, as literature considers them as mood rather than emotions [35]. Emotions usually influence physiology only for a short period [27].

We ruled out articles utilizing only EEG signals, as devices suitable for everyday life are yet to be developed. Current solutions are reliable only in stationary position, as any movement can cause artifacts in recorded brainwaves due to: (1) change of the position of electrodes, (2) activation of other brain areas, (3) electrical activity of muscles. Additionally, due to their design EEG headbands may be uncomfortable to wear for longer periods of time.

We also excluded articles classifying only a single emotion, e.g., only *fear*, as they do not prove that recognizing multiple emotions is really possible. Studies focusing exclusively on arousal were also excluded since arousal by itself

does not represent emotions. Recognition of valence levels was an exception from this rule because it can distinguish positive and negative emotions, e.g., happy vs. sad. Some papers examined emotional states but did not explicitly carry out emotion recognition. Instead, they performed statistical analysis to find correlations, e.g., [36]. We decided to exclude them as they did not prove that emotion recognition is possible.

The quantitative results of the SLR are in Supp. Mat.

## 4 General Study Design for Emotion Recognition

### 4.1 Lab vs. Field Studies

Planning the study in a laboratory setup differs from planning it in an uncontrolled environment. Although the general components are similar (Fig. 2), the details and implementation of each component may vary significantly, see Tab. 2. In principle, we can distinguish nine components in research design for emotion recognition.

A decision about the emotional model is crucial because it influences other components, especially the stimuli and self-assessment preparation, as well as the reasoning output. Although all current models are applicable to both lab and field studies, the complex ones complicate the research. For example, it is easier to find stimuli eliciting two different emotions - happy vs. sad, than to find stimuli evoking eight different emotions. In the case of the latter, several emotions are likely to co-occur for one stimulus. The second stage covers recruitment, profiling, selection, and training of study participants – subjects. Some of them need to be excluded due to diseases, e.g., heart problems can interfere with the electrocardiographic (ECG) or blood volume pulse (BVP) signals. A simple explanation of the test procedure followed by the sample questionnaire is usually sufficient to prepare the subjects for the lab study. The researcher may attach and calibrate all the devices themselves and assist the subject throughout the test. Contrarily, training the subjects for the field study should include information on how to put on and care for the device, as participants themselves will handle the devices every day. Technical support should also be provided, preferably 24/7.

For laboratory studies, some stimuli to elicit emotions should be prepared and selected. They can be carefully planned, consumed without distractions, controlled, and interrupted if necessary. They are also known in advance and are served to all subjects in the same way. However, the range of emotions is then limited to these selected stimuli. On the other hand, field studies do not require any stimuli preparation because unpredictable real-life events evoke emotions. Emotions experienced in such way occur in their natural context and can be much more intense and rich. The real-life stimuli will be different for each subject and usually unknown to the researchers.

Next, physiological signals from wearables are traced. The laboratory setup enables measurement with multiple, medical-level devices providing high-quality signals. A usually stationary body position limits the number of possible artifacts. Devices used in the field study have to be wearable and convenient. It results in lower quality of sensors and signals [37], making measurements laden with inaccuracies and artifacts. Field studies are also more engaging for participants as they, e.g., need to charge and sync the devices.

Self-assessments have to be collected together with physiological signals. Questionnaires in the lab can be triggered and filled out right after the stimuli consumption. They can be detailed and precisely designed to match the selected stimuli. Self-assessment in the real-life scenario, in turn, is a vital challenge. It is nearly impossible to accurately determine when subjects experience emotions, thus, when the self-assessment should be triggered. Although, our recent idea employing a pre-trained model may help to address this issue [38], [39]. Most often, the self-assessment is triggered at random or on demand. The questionnaire should be as short as possible, and even then, it might be troublesome to fill it out immediately after experiencing emotions [26].

Later, raw signals are preprocessed, sampled, synchronized, and descriptive features are derived. The selected features combined with the ground truth emotional labels are used to train the reasoning model. In order to adjust the model and make it ready for real-life application, hyperparameter optimization should be considered. The reasoning models trained on data collected in the lab may perform well on other data from another lab study; however, they will probably be useless for applications outside the sterile laboratory setup. The models developed in the field may demonstrate slightly lower accuracy, but they are more suited to real-life scenarios.

### 4.2 Emotion Recognition Scenarios

Based on the papers from our SLR, we can identify seven main scenarios for emotion recognition, Fig. 3. The first five refer to lab studies, whereas the last two to research in the field. They mainly differ in the approach to (1) labeling, i.e., how the emotional class label, which defines ground truth, is obtained, and (2) stimuli, i.e., what elicits the emotions. Scenario 1 was exploited in [22]–[24], [40]–[45], and probably in [46]?, [47]? ('?' denotes it was deduced by us); scenario 2 in [42], [48]–[52], [53]?; 3 in [54]–[59], [46]?; 4 - [60]; 5 - [31], [47]?; 6 - [20], [61]–[63], and scenario 7 in [21], [26], [64], [65]? We were not able to identify scenario in [19]. Further details, including emotional models converted to machine learning problems along with references to individual papers, can be found in Supp. Mat. Tab. 5.

## 5 Crucial Research Components

### 5.1 Emotional Models

The key assumption of emotion recognition from physiological signals is the belief that there is a relationship between the subjective and conscious recognition of experienced emotions and their objective manifestation in physiological signals. This subjective recognition can be performed by subjects themselves, their peers or external experts, as well as by researchers planning the study and assigning labels to the stimuli, Fig. 1. They are further used as the ground truth while training machine learning models, Fig. 3 and 5. Hence, the labeling process is commonly based on a predefined, fixed list or dimensions of emotions investigated in a given study. To create such a list, the researchers made use of psychological achievements in the field. None of the studies in our SLR utilized free-text descriptions that would enable them to identify new types of emotions, see, e.g., [66].

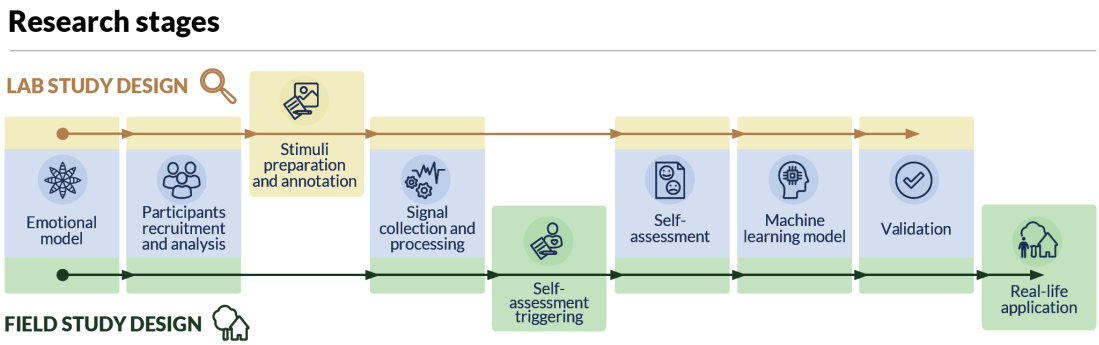


Fig. 2. Common and unique research stages (study design) for emotion recognition in the lab and in the field.

TABLE 2
The main differences in emotion recognition between lab study and field study. '+' denotes an advantage; '−' is a disadvantage; '±' means an aspect has both, positive and negative sides

| Category | Lab study | Field study |
|---|---|---|
| Emotions experienced | − In controlled environment<br>− Impacted by unnatural conditions<br>− Limited to the prepared stimuli<br>+ Beginning and end determined by the stimuli | + In natural context<br>+ Full range of emotions<br>− Occurrence is difficult to capture<br>− Hard to determine the beginning and end |
| Stimuli | ± Planned and prepared, e.g., videos, images, music, tasks<br>+ Fully controlled by researchers, may be interrupted<br>+ May be annotated<br>+ Known duration<br>+ No distractions nor unexpected stimuli<br>+ Condensed sequence of stimulants separated by wash out | + Daily life stimuli<br>− Unknown stimuli<br>− No stimuli label<br>− No starting point<br>− Unknown duration<br>− Out of researcher's control<br>− Susceptible to life conditions, e.g., drugs, fatigue |
| Labeling (ground truth) | + Self-assessment<br>+ Expert-annotated stimuli<br>+ Observed and derived by external experts | − Mainly self-assessment<br>+ Nearby person (relative, friend) |
| Self-assessment | + Detailed<br>+ Often<br>+ Trigger time easy to determine<br>+ Triggered and filled out right after each stimuli | − Limited scope<br>− Sporadic<br>− Triggering time is difficult to determine<br>± Self-, event-, activity-, randomly-triggered, schedule, reasoning<br>− Usually delayed participant's response [26] |
| Measuring physiology / devices | + Medical-level, precise devices<br>+ Devices can be large and wired<br>+ Many devices simultaneously possible<br>+ External devices possible, e.g., multiple cameras<br>+ No battery problem<br>− Stressful condition<br>+ High-quality signal / data (little external interference)<br>+ Stationary position (usually sitting) | − Lower quality of sensors and signals [37]<br>+ Personal, convenient, useful wearables<br>− Only few devices feasible<br>± Battery-efficient wearables<br>+ Convenient and unnoticeable measuring<br>− Artifacts caused by the movement and field conditions<br>± Data transfer to server (in real-time / post-session)<br>− Lack of data when wearable is off / not worn<br>− 24/7 technical support required |
| Additional factors | + Static environment (temperature, lighting, etc.)<br>+ Meta-questions (e.g., health issues, time past since last coffee/activity/sleep)<br>± Relatively small amount of data | − Variable environment<br>− No meta-question<br>± Large amount of data to be collected and processed |

Since human emotions are complex, there is no single, commonly agreed emotional model. Many models have been developed by psychologists for decades, e.g., the first dimensional model from 1954 [28], or four *primary* emotions distinguished by Krech et al. in 1974 [47], [67]. The most frequently referenced models were proposed in the 1970s and 1980s. Ekman and Friesen identified six emotions recognizable from facial expressions: happiness, anger, fear, sadness, disgust, surprise [68]. They actually followed Darwin's findings 100 years before [69]. Plutchik proposed eight basic bipolar emotions: joy vs. sadness; anger vs. fear; trust vs. disgust; and surprise vs. anticipation [70].

Overall, we can distinguish two general models of emotions: (1) **Discrete** emotional models that treat emotions as a set of distinct categories, e.g., Ekman's, Plutchik's, Izard's ones; and (2) **Dimensional** emotional models, e.g., Russell's circumplex [71], PAD (Pleasure-Arousal-Dominance) equivalent to valence-arousal-dominance model used in the very popular SAM (Self-Assessment Manikin) [72]. Izard claims that both discrete and dimensional approaches to emotions may complement each other [73]. There are also some discrete approaches, which combine multiple affective states, including stress or anxiety [5], [24].

We can synthesize issues related to emotional models with the following general remarks:

1) Most researchers (76% papers) use their own emotional

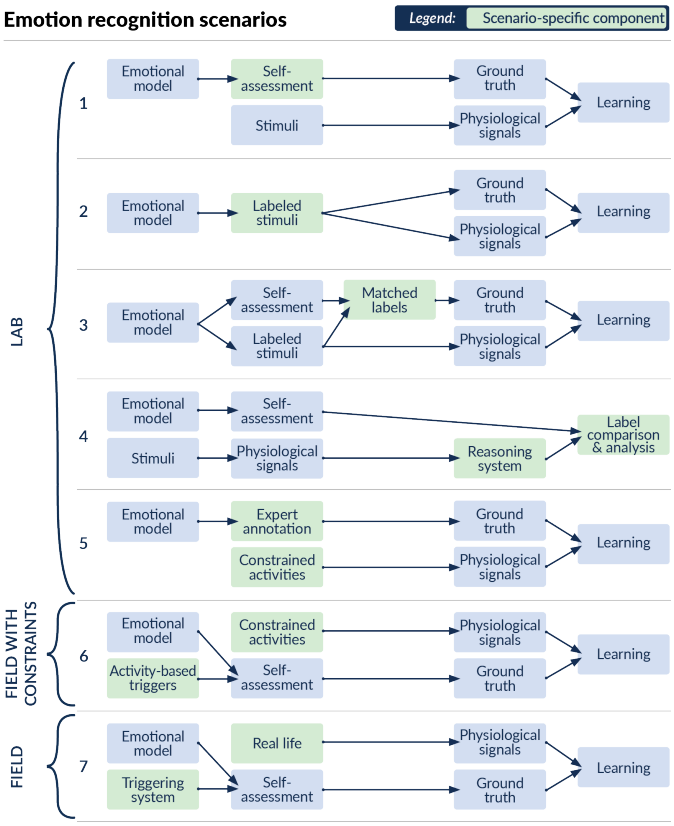

Fig. 3. Emotion recognition scenarios identified in SLR.

models, usually discrete ones (65%). Psychological models were usually just an inspiration. They are significantly transformed rather than directly exploited. This prevents any kind of comparison and research replicability.

2) Transition from psychological models to ML tasks usually loses achievements of psychology, especially relationships between individual emotions.
3) Dimensional emotional models are often converted into binary or multiclass ML problems (only $^{2}/_{16} = 13\%$ regressions). This makes them actually discrete ones. Such transition loses information about relations between emotions (classes): *low valence=sad* is as distant from *neutral* as it is from *high valence=happy*. This is a common problem of converting numerical values into categories.
4) Dimensional emotional models assume independence between dimensions, so they should be implemented with independent ML tasks. It was respected in $^{8}/_{11} = 73\%$ papers.
5) Multiclass classification is based on discrete emotions, which cannot co-occur.
6) The meaning of some discrete emotions is unclear. For example, Lisetti and Nasoz used two discrete emotions [48], [49]: *anger* and *frustration*. Psychologists claim that "frustration consists of a subset of the attributes of anger" and "anger is a particular kind of frustration and thus frustration is a separable part of anger" [74].
7) Some discrete emotional models are almost the same even though the researchers distinguish them. For example, Rattanyu et al. [54], [55] stated they were inspired by Plutchik and Circumplex and Nguyen et al. [64] started from Oatley-Johnson [75]. Both groups finally used the same five discrete emotions that correspond to the Ekman-Friesen model but without *surprise* [68].
8) The list (or dimensions) of possible emotions is usually fixed in advance at the early stage of study design. The participants cannot add and describe their own emotional states. Therefore, they have to tailor their perception to the pre-defined borders.

### 5.2 Stimuli

Depending on the environment in which studies were performed, the researchers had to face various issues with emotion elicitation, Tab. 2. To approach them, different types of stimuli were employed. They can be grouped into four categories, two per environment.

In a classical laboratory setup, stimuli consist of simple activities like watching scenes from movies, giving researchers complete control over the process of emotion elicitation. Moving towards everyday life, researchers sacrifice some knowledge about the stimulus to gain more real-life emotions. In field studies, researchers have no information about the background of emotion, but they obtain more natural experiences and corresponding physiological signals.

In all types of studies, subjects' physiological responses can be influenced by a set of factors like sleep quality or the use of substances affecting physiological reaction, e.g., coffee or drugs. In lab studies, it is easier to exclude such participants or take these factors into account. In this way, any potential bias in stimuli perception and reaction to them can be filtered. Such filtration in field studies is hardly possible due to little information about the factors and no information about stimuli, Tab. 2.

Field studies to recognize real-life emotions were carried out in eight papers (24%). Such a setup allows measurements of emotions during unexpected situations. It comes at the cost of decreased knowledge about the stimulus, its duration, and the time that passed between stimulation and self-report. Scenarios in the field can be divided into two categories: (1) with some set of limitations applied, or (2) without any constraints, Tab. 3. In field studies with constraints, the amount of possible stimuli decreases, which is useful when researchers want to focus on a particular setup or exclude the impact of some factors. The following study limitations were specified in five papers: walking the particular route in the city [61], [62] or around one city park [63], performing classroom activities [20], or working in the factory [65]. If researchers want to focus on the subjects' whole daily lives, they should implement field studies without any limitations (real-life scenario), which was done in four included papers. In one paper [47], the authors did not specify the environment where experiments were conducted, nor the type of stimuli used.

As we expect the studies to move from the lab to the field, it would be beneficial to explore and compensate for the context in which emotional experience emerges. Various factors like coffee or medication, which are sometimes important components of human lives, can affect the reactions to stimuli and the traced physiological signals.

### 5.3 Context Impacting Emotion Experience

The context can potentially affect the process of data acquisition. We distinguish two categories of context affecting either subjects' physiology or their perception. The latter is important for self-assessment questionnaires. Gathering

TABLE 3
Stimuli in field studies

| Environment | Stimulus | Used by |
|---|---|---|
| Field with constraints | Walking on specified route | [61], [62] |
| | Taking photos in the park | [63] |
| | Working in a factory | [65] |
| | Classroom activities | [20] |
| Field | Real life | [21], [26], [64] |

data on the context in field studies is difficult. Subjects may have trouble accurately estimating the time that has passed since their last cup of coffee, be busy with other activities, or forget to submit relevant data. Only two papers collected additional information on the temporary context in field studies. Although Schmidt et al. [21] stored data on physical activities and sleep quality, they did not take this data into account while reasoning. It is unclear whether Dao et al. [20] utilized data on physical activities in the emotion recognition process. They stated that activities were used to uncover lifestyle-mood patterns of subjects and that some patterns were associated with emotions. Hu et al. [47] and Majumder et al. [53] imposed context on participants instead of surveying it. They monitored signals during resting (sitting) and active (walking) states. Even though their papers lacked a comparison of these approaches, it may be a good idea to perform such studies and find differences between emotions in various contexts.

Although 91% papers included some information about the long-lasting factors influencing emotions like age, gender, health condition, only about a fourth of studies documented temporary factors like consumption of drugs. No research took them into account during analysis. Questions and tests about different factors were rather used to ensure that no participant was under the influence of any substances. Even if the context was tracked, it was not used to compensate in any way for differences caused in physiological signals. Schmidt et al. [21] stated that they did not consider context as it is rarely available. It is also unclear whether the context supported the emotion recognition process in the experiment by Dao et al. [20]. A more extensive discussion of the context is in Supp. Mat. Sec. 3.3.

### 5.4 Emotion Self-assessment

Simple labeling of subjects' emotions with labels assigned to stimuli can be done easily in the laboratory, where each stimulus is designed to elicit a particular reaction. However, such an approach is impossible in field studies where researchers cannot control either the type or duration of the stimuli. Besides, stimuli-based labeling does not respect the subject's subjective perception, i.e., a given stimulus can evoke no, or completely different emotions than assumed.

To tackle these issues, researchers employed self-assessment questionnaires. This type of labeling requires subjects to report their feelings themselves, most frequently within some pre-defined categories. Among articles included in our SLR, 79% used self-assessment, most commonly the well-known methods or their modified versions. These included Self-Assessment Manikin (SAM) [72], [76], AniSAM and AniAvatar [77], PANAS [78], Multidimensional Mood Questionnaire (MDMQ) [79], or Ecological Momentary Assessment (EMA) [80], [81]. Some authors (12 papers, 35%) developed their own questionnaires, which included rating the intensity of emotions or affect on a Likert scale [56], [57], with a continuous scale [22], by choosing emotions from available categories [20], [64] or from binary values [41], or by means of keyword descriptions [23].

Several different emotional models and questionnaires are used in self-assessments. As both dimensional and discrete emotional models attempt to describe affect, one can be translated to another. Such translation was performed in eight papers. In papers utilizing self-assessments, the final models of emotion were discrete in 16 papers and dimensional in 11 papers, Tab. 4.

Self-assessment questionnaires may be hard to use for subjects who have not been trained properly. It is especially important when emotions are described with vague, ambiguous, or not intuitive terms, e.g., *dominance*. It can be unclear whether high value means that we want to dominate someone after experiencing emotions or the emotion is dominating us. Even the trained subjects may get lost in what different levels mean in stressful situations.

Papers describing the use of self-assessment questionnaires rarely provided any details on the interface with which subjects interacted. Martens et al. [22] had the subjects rate their emotions on five sliding scales with ticks marked for orientation and borders described in words. A similar design was applied by Exler et al. [26]. Kim et al. [63] utilized three-dimensional SAM for pleasure, arousal, and dominance in their application. Additionally, subjects rated the memorability of a taken photo on a separate sliding scale. Kanjo et al. [61] had subjects rate valence using SAM with five possible values. All of these papers contained screenshots of their applications. Wampfler et al. [45] provided an interface screenshot with math tasks, but not the one with emotion ratings. Albraikan et al. [42] mentioned utilizing AniAvatar in self-reports, but they did not describe them in more detail. In their other paper [60], we can find screenshots of the application for providing subjects with feedback on their emotions using emojis. However, they did not specify the form of questionnaires given to subjects. Nevertheless, these two approaches may be a good way of representing emotions in a straightforward fashion.

As stated in Supp. Mat. Sec. 3.3, no paper reported taking the context into account. In lab studies, substances that influence emotional experience may be considered by asking subjects to fill in special forms or not to use any such substances [41]. However, forbidding subjects from, e.g., taking medication or exercising would be inconvenient during field studies, so in our opinion, self-assessment questionnaires should include questions on the context. They would allow researchers to include adequate information in analysis and inference. To our best knowledge, only in two field studies gathering context was considered. Schmidt et al. [21] declared that they did not use additional information during emotion classification. Dao et al. [20] did not describe their process in much detail, so it is not clear whether the context was supplied to the model for emotion recognition.

In field studies, researchers have to face not only limited or no information about the stimulus, but they also need to consider the usability of the employed system. Depending on the setup, different ways of triggering self-

TABLE 4
Trigger time and type of self-assessment, as well as emotional model utilized in studies; Dimen. – dimensional.

| Environment | Trigger | Questionnaire | Model of emotions | Used by |
|---|---|---|---|---|
| Lab | After stimulus | SAM | Dimen. | [43]–[45] |
| | | SAM | Discrete | [40], [54], [55] |
| | | AniAvatar | Discrete | [42], [60] |
| | | PANAS | Discrete | [58], [59] |
| | | Own | Discrete | [24], [46], [48], [49], [56], [57] |
| | | Own | Dimen. | [23], [41] |
| | Time dependent | Own | Dimen. | [22] |
| Constrained field | Quasi continuous | SAM | Dimen. | [61], [62] |
| | After taking a photo | SAM | Dimen. | [63] |
| | Voluntary | Own | Discrete | [20] |
| | No info | Own? | Discrete | [65] |
| Field | Voluntary, randomly, EMA | SAM | Dimen. | [21] |
| | On events, time dependent, voluntary | MDMQ | Dimen. | [26] |
| | No info | Own? | Discrete | [64] |

assessments were used, Supp. Mat. Fig. 4. For the field studies in constrained environments, researchers prompted the subjects: (1) after taking a photo with a phone [63], or (2) quasi-continuously [61], [62]. The second approach is suitable only for studies lasting for a short period of time, as subjects in [20] reported inconvenience in continuously tagging emotions during lessons. Additionally, in this paper, the authors built a system able to recognize the emotions of subjects, who later voluntarily self-validated its predictions.

For collecting emotions during daily activities, researchers employed different, often less demanding questioning strategies. They were based on the following methods of self-assessment triggering: (1) voluntary [21], [26], [64]; (2) at the fixed time, e.g., every full hour [26]; (3) event-based, e.g., when subject received a message, ended a call, or after a meeting scheduled in the calendar [26]; and (4) invoked randomly [21]. In Supp. Mat. Fig. 4, we also include a new concept to employ some decision model to trigger self-assessment, which was proposed in [38], [39]. Such a model could detect possible emotional experience and make a decision based on the subject's recent physiology. This solution would allow researchers to enhance the process of data gathering, potentially increasing the number of annotated cases for each emotion.

In almost all papers describing field studies (88%), self-assessment questionnaires were prompted on a phone. Kadoya et al. [65] did not specify what type of questionnaires were shown to participants. We deduce that either a phone or watch would be the most suitable for prompting questionnaires in their case. Exler et al. [26] utilized questionnaires presented on both a phone and a watch.

### 5.5 Biosignals and Sensors

The human body is a complex, dynamic and nonlinear system with a vast number of feedbacks between individual organs and entire physiological systems, which simultaneously ensure internal homeostasis as well as an appropriate response to changes in an external environment, Fig. 4. External stimuli excite the central nervous system (CNS) through the sense organs and are analyzed primarily by the cerebral cortex. This activity is reflected in the electroencephalogram, typically recorded by 2-20 skull electrodes, including the passive one. The main features of the EEG are the brainwaves with energy varying in the delta, theta, alpha, beta, and gamma subbands. Moreover, the previous states of the cortex, including memories, also affect its current behavior. The CNS mainly controls skeletal muscles, also counting the respiratory ones. A specific muscle's electrical reaction can be measured as the electromyogram (EMG), and any selected body part's movement can be measured via accelerometers (ACC). These are helpful in recognition of the subject's physical activity. This muscle activity is often seen as artifacts when other biosignals, especially electrical, are monitored. The specific effects of respiratory muscle action are chest and abdomen movements, as well as airflow ventilating the lung. Both represent the depth and rate of breathing, defining the respiratory waveform (Resp). Gas exchange between the alveoli and blood results primarily in the level of oxygen saturation and blood pH, monitored by chemoreceptors and then analyzed by the autonomic nervous system (ANS) cooperating closely with the central one. In reaction, the electrical stimulation of heart activity is regulated, triggering the depolarization and repolarization of the heart atria and ventricles. These phases are well visible in the electrocardiogram (ECG) waveform, dominated by the QRS complex representing the depolarization of the ventricles. The heart stroke volume and rate determine the pulsatile blood flow, and together with arterial smooth muscle tone (controlled by the ANS), also temporary blood pressure (sensed by baroreceptors) - effects monitored together noninvasively as the photoplethysmogram (PPG) or blood volume pulse (BVP) signal. The ANS also controls the tone of airway smooth muscle modifying respiratory mechanics, that may lead, e.g., to abrupt panic-induced breathlessness. Simultaneously, the intensity of blood flow affects body temperature regulation and determines the speed of transport of hormones produced by the secretion system. The effects of both processes are reflected, among others, in the measured body (BT) or skin temperature (SKT), and skin sweat glands secretion modifying the electrodermal activity (EDA), also known as galvanic skin response or resistance (GSR), electrodermal response (EDR), or skin conductance (SC). This signal has two visible components: slower tonic skin conductance level (SCL) and faster phasic skin conductance response (SCR).

The SLR has revealed that the following signals are primarily used for affective computing, Tab. 5: EEG, ECG, and EMG produced by the electrical activity of the brain, heart, and skeletal muscle respectively; EDA (also called GSR, EDR, or SC) representing electrical properties of the skin; pulsation of arterial blood measured as PPG or BVP, body temperature denoted BT or SKT, respiratory waveform Resp, and movements acceleration ACC, which is often supplemented by GPS position or measurements done with gyroscope (GYR).

The above signals are recorded directly applying dedi-

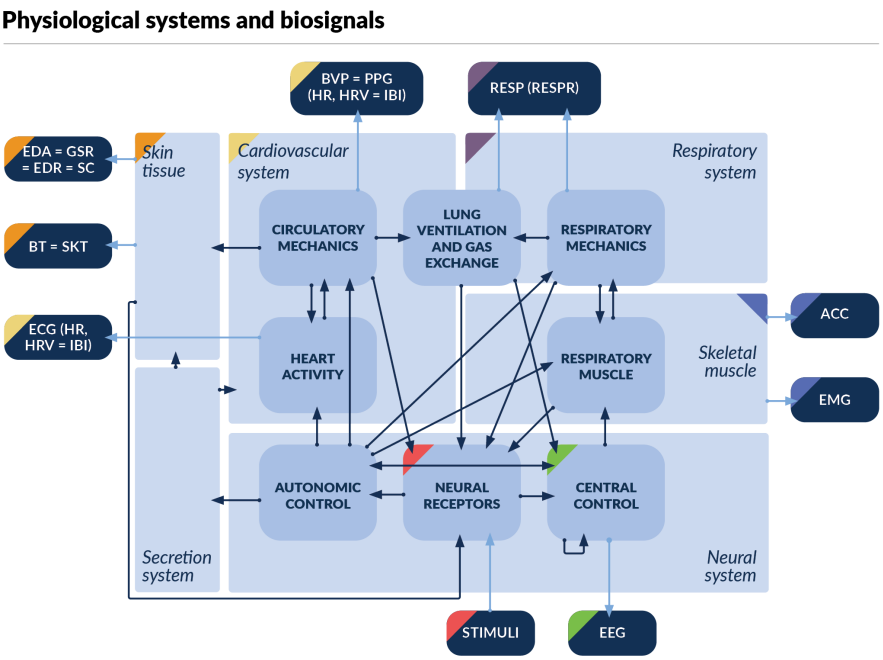


Fig. 4. Interrelationships between physiological systems and biosignals.

TABLE 5
Physiological signals used for emotion recognition

| Signal | Characteristics | Used by |
|---|---|---|
| EEG | Amplitude: 5-300 μV; Brainwaves: delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-14 Hz), beta (14-30 Hz), gamma (30-80 Hz) | [22], [23] |
| ECG | Amplitude: 0.5-5 mV; Frequency spectrum: 0.05-150 Hz | [26], [47], [54]–[57] |
| EMG | Amplitude: 0.1-5 mV; Frequency spectrum: 20-2500 Hz | [21], [44] |
| EDA, GSR, EDR, SC | Amplitude: 1-20 μS; Tonic component (SCL) and Phasic component (SCR) | [19]–[22], [24], [31], [40]–[46], [48], [49], [52], [53], [60]–[63] |
| PPG, BVP | Amplitude: 20-300 mmHg; Frequency spectrum: 0-15 Hz | [20], [21], [23], [24], [41], [43], [44] |
| BT, SKT | 35-42 °C | [20], [21], [40]–[42], [44]–[46], [48], [49], [53], [60]–[62] |
| Resp | Frequency spectrum: 0-5 Hz | [21] |
| ACC | Range: 0-2 g | [20], [21], [58], [59], [61]–[63] |

TABLE 6
Physiological signals' derivatives

| Signal | Source signal | Characteristics | Used by |
|---|---|---|---|
| HRV, IBI | ECG, PPG, BVP | Amplitude: 0.3-1.5 ms; Very low frequency component: 0.004-0.04 Hz; Low frequency component: 0.04-0.15 Hz; High frequency component: 0.15-0.4 Hz | [19]–[21], [24], [41], [42], [44], [45], [52], [54]–[57], [60] |
| HR | ECG, PPG, BVP | Range: 40-210 bpm | [20], [21], [24], [26], [40], [42], [43], [45], [46], [48]–[51], [53]–[55], [58]–[65] |
| SCL, SCR | EDA (GSR, EDR, SC) | SCL spectrum: 0-0.16 Hz, SCR spectrum: 0.16-2.1 Hz | [21], [41], [45] |
| RespR | Resp | Range: 0-40 breaths pm | [21], [44] |

cated sensors, but other biosignals useful in affective computing can be derived from them using special algorithms, Tab. 6. The most informative one is heart rate variability (HRV), describing changes in the duration of interbeat intervals (IBI). IBI are typically calculated as the intervals between R peaks in the QRS complexes of ECG (the best visible peaks) or by finding the peaks in the PPG/BVP signal (alternatively, a zero-crossing method with hysteresis can be used). The HRV waveform reflects plenty of feedbacks in the ANS controlling the heart rate (HR) and stroke volume (SV), which are related, e.g., to gas, glucose and hormone transport, acid-base balance, body temperature regulation, respiration efficiency, or skin glands secretion. Since SV modulation is only moderate, HR plays a key role as a controlled variable. It is measured with the number of heartbeats (peaks identified in ECG or PPG) per minute (bpm). Similarly (by finding peaks or applying the analysis of frequency), the Resp waveform can be recalculated into the respiratory rate (RespR), representing the number of breaths per minute.

To summarize, the external and internal stimuli influence the affective state of a human, and thus also the state of physiological systems via a net of biofeedbacks, having a more or less delayed impact on all the biosignals mentioned above. The phasic skin conductance response to the stimuli is usually monitored for one to four seconds after the stimuli onset [82]. As a result of aroused state, the SCR amplitude gradually increases and achieves the peak, after which the recovery phase begins - a decline in the phasic component. Depending on where the SCR is monitored, the peak may be visible sooner or later. In [83], SCR response-time on various stimuli types was analyzed. The reaction occurred within 1-5 seconds after the stimuli onset and reached the peak after 3.9 s (measured on palm), 4.3 s (finger), and 5.0 s (foot). In [84], the SCR peak occurred within 3-4 seconds after the stimuli onset. Contrary to the SCR, the body response

measured as change in brain activity, heart rate variability, breathing rate, or muscle activity can be observed much faster, usually right after the stimuli onset.

### 5.6 Wearables

There are countless devices for monitoring and measuring physiology. For lab studies, medical-level, precise devices can be utilized, such as BioPac MP160 or ProComp Infiniti. They can be large, wired, and sophisticated - after all, a trained technician or researcher will operate them.

For field studies, however, small and easy-to-use wearables, which the participants can handle, are necessary. The wearable device should be selected according to the requirements and needs of a given study. Wearables are of different sizes and can be worn on various parts of the body. On the market, we can find smartwatches, smart bands, wristbands, fit bands, armbands, headbands, chest straps, chest patches, smart rings, smart glasses, smart clothes, and more. While selecting a device for the field study, the following crucial issues should be considered:

- availability of the physiological signal in a raw format;
- signal frequency suitable for the research problem;
- signal quality – free from artifacts, obtained with a well-calibrated and adhering to the skin sensor;
- portability;
- ensuring data synchronization, e.g., data transfer to the cloud, integration with existing study system;
- convenience - ease of configuration and use (putting on/off, charging), usability for the user, battery life.

As part of the SLR, we reviewed more than 50 wearable devices available on the market in terms of sensors, raw signal availability, and other data helpful in emotion recognition. For a detailed analysis and discussion, please refer to [85] and [16].

Tab. 7 contains the devices most frequently chosen in the SLR articles. The most popular was the relatively old wristband Empatica E4, launched in 2015. Its advantages include several sensors, high frequency, raw signal availability, battery lasting about 30 hours of recording, and the possibility to integrate through a dedicated API. Its measurements in the non-movement setup are comparable to the ambulatory monitoring system but are prone to artifacts and useless when the subject is in motion [86]. Its downsides are very high cost, lack of screen, which sometimes causes uncertainty about the device's current state, and no additional feature useful for participants. Taking into account all aspects (especially the artifacts in motion), Empatica E4 is not a good candidate for daily data collection. Moreover, Borrego et al. [87] showed that Empatica's EDA measurement in response to emotionally-valenced images is much worse than with a laboratory-grade device.

Another popular device – Microsoft Band 2 smartband, is even older. It offers similar sensors but also many functionalities for the participant: activity and sleep tracking, integration with a smartphone, watch functions, etc. It was also found to be precise for stationary measurements [88]. The affordable EEG headbands, e.g., Emotiv Insight, provide a great opportunity for precise emotion recognition, but they can only be used in certain, non-movement setup, such as in the lab or at home. The EEG headbands are not suitable for monitoring emotions all day long because (1) they record artifacts from all around, and (2) it may be uncomfortable to wear them for so long.

Other devices mentioned in the included articles are: BodyMedia SenseWear; chest strap Polar H7 and smartwach Samsung Gear 2 [58], [59]; chest strap Polar H10 [45]; RF-ECG biosensor kit [54], [55]; XYZlife Bio-Clothing [56], [57]; Q-sensors [31]; Biopac BioNomadix MP150 [43]; chest strap ekgMove [26]; wristband Silmee W20 [65]; smartwatch Algoband F8 [51]; wristband Mio Link [64]; self made smart clothing used in [47]; Wacom Bamboo Ink stylus, Shimmer GSR+, and PPG ring [45].

In our opinion, smartwatches will gain in popularity in the coming years, as they are pervasive, cheap, very useful to the user, offer various sensors (BVP, ACC, GYRO, and sometimes ECG), and allow for custom made applications. Usually, the lack of the EDA sensor is the only shortcoming in applying smartwatches in emotion recognition field studies.

### 5.7 Machine Learning Procedure

In the most common approach to emotion recognition, the collected raw signals are preprocessed, sampled, synchronized, and descriptive features are derived. Next, simple classifiers or deep neural networks are iteratively trained and optimized to achieve the best possible predictive model, Fig. 5. We call it the *classical feature-based* approach to emotion recognition. It requires domain-specific, expert knowledge about the sensors and signals to extract meaningful and informative features. Alternatively, an *end-to-end* concept can be used, which omits signal preprocessing and feature extraction, i.e., the acquired raw signals are directly passed to the deep learning architectures assuming they will be able to extract the essential information on their own. See Fig. 5 for an illustrative comparison between these two approaches to the emotion recognition problem. The end-to-end approach is a recent idea and has been explored very little, but it is a promising direction [21], [89], [90]. In the reviewed papers, the authors tend to apply the classical feature-based approach.

Tab. 8 summarizes frequent approaches to various machine learning stages. The majority of works (88%) utilized classic classifiers, such as decision tree (DT), k-nearest neighbors algorithm (KNN), support vector machine (SVM), simple Neural Networks, to solve a multiclass problem. Deep neural networks (convolutional, CNN; long short-term memory, LSTM) appeared in only four papers. They also commonly did not consider the data imbalance problem (88% of papers), at the same time applying the accuracy as a quality measure (82% of papers). Such combination is very arguable, as accuracy may achieve high value due to the classifier focusing on the majority class. Furthermore, only few authors (24%) applied statistical tests on the results. For details for each included paper, see Supp. Mat. Tab. 6.

### 5.8 Transition from Psychological Models to ML Tasks

Transition from the psychological models to a simple set of discrete emotions may lose some information about mutual relationships between component emotions. Note that psychological models often provide sophisticated dependencies between emotions, e.g., position of the emotional state

TABLE 7
The most popular devices in SLR that measure physiology

| Device | Type | Release date | Sensors | Physiological raw signals | Other data | Used by |
|---|---|---|---|---|---|---|
| Emotiv Insight | Headband | 2015.10 | EEG, ACC, GYRO, MAG | EEG | ACC, GYRO, MAG | [22], [23] |
| Empatica E4 | Wristband | 2015 | PPG, EDA, ACC, TERM | BVP, EDA, SKT | HR, PPI, ACC, tags | [20], [21], [23], [24], [40]–[43], [45], [50], [53], [60] |
| Microsoft Band 2 | Smartband | 2014.10 | PPG, EDA, ACC, GYRO, TERM, BAR, ALT, AL, UV | BVP, EDA, SKT | HR, PPI, ACC, GYRO, BAR, ALT, AL, STP, CAL, UV | [19], [22], [44], [52], [61]–[63] |
| BodyMedia SenseWear | Armband | 2003 | EDA, ACC, TERM | EDA, SKT | ACC | [46], [48], [49] |

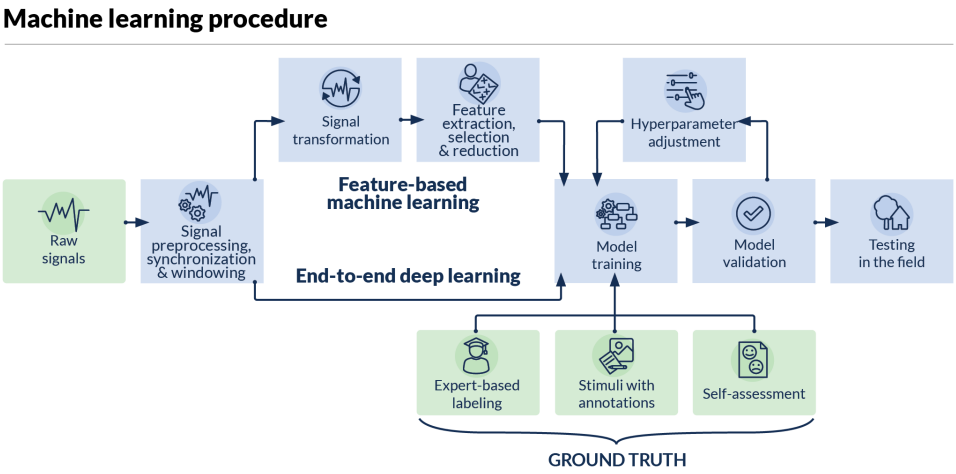


Fig. 5. Machine learning process for emotion recognition using biological signals from wearables. Blue blocks correspond to individual stages. Green blocks indicate input – signals and ground truth.

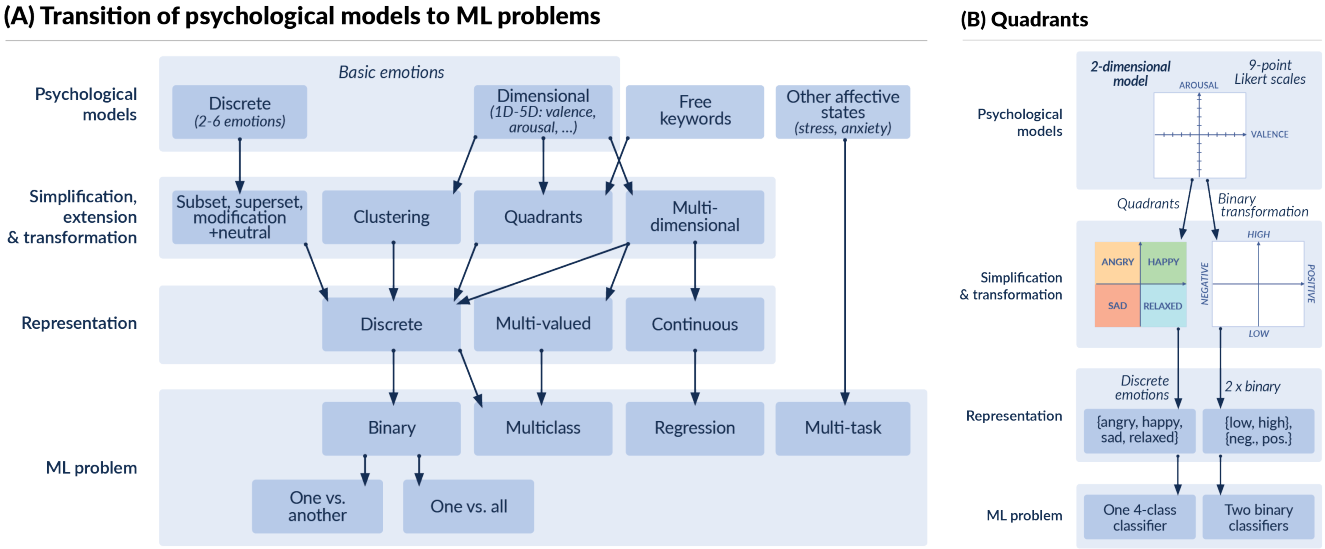


Fig. 6. **(A)** Transition of original psychological emotional models into machine learning used in experiments. **(B)** The 2-dimensional emotional space *arousal-valence* converted to four quadrants, next exploited as (1) 4-class classification or (2) two binary models, one for each dimension.

and its angle towards other emotions matters in Russel's Circumplex. In Plutchik, in turn, *fear* is located opposite *anger*, *joy* is close to *anticipation* and *trust*, but far away from *sadness*. All this information developed by psychology is missing if used in the form of simple binary or multiclass classification, in which discrete emotions are just treated as an unordered set of distinct elements, Fig. 6, Supp. Mat. Tab. 5. Then, every discrete emotion is equidistant from all others.

Similarly, multidimensional models are also converted to discrete values a lot, e.g., using quadrants, which later on are treated as independent four values. This directly leads to 4-class multiclass classification model [23], [41], [50], Fig. 6. This obviously refers to 1-dimensional initial models, which distinct values were direct output classes [26], [61], [62].

Nevertheless, some authors respected the initial independence assumption between individual dimensions and trained a separate model for each dimension. Then, binary or multiple values of each dimension were the classifier output. It resulted in two binary models [19], [41], [43], [44], one for each dimension; or two 3-class models [21], [42]. Independently, having multiple discrete emotions collected, the researchers considered binary machine learning models one-against-all repeated as many times as the number of

TABLE 8
Approaches, methods, and measures used at particular machine learning stage. '?' means it was not described but inferred by us only

| Stage | Approach | Used by |
|---|---|---|
| Classification type | multiclass | [21], [23], [24], [26], [31], [40]–[42], [45]–[49], [51], [53], [56], [58]–[62], [54]?, [55]?, [64]?, [20]?, [50]?, [52]?, [65]? |
| | binary | [31], [41], [43], [44], [57]–[59], [19]? |
| | regression | [22], [24], [63] |
| ML models | classical (KNN, SVM, etc.) | [19], [21], [23], [24], [26], [31], [40], [41], [43]–[59], [61], [64] |
| | deep (CNN, LSTM, etc.) | [21], [23], [59], [62] |
| Quality measures | accuracy | [19], [23], [24], [26], [31], [40], [42]–[62], [64] |
| | F-measure | [21], [42], [44], [58], [61], [62] |
| | Other | [40] – conf. level; [31] – AUC, precision, recall; [41] – correct classif. ratio; [62] – precision, recall, error rate, RMSE, confusion matrix; [44] – ROC curves, confusion matrices; [45] – micro-avg AUC, macro-avg AUC; [22] – MAE, RMSE, Pearson correlation coefficient |
| Imbalance in learning samples | not considered | [19], [20], [22]–[24], [26], [31], [40]–[44], [46]–[56], [58]–[65] |
| | considered | [57] – equal size sampling; [45] – RF with balanced class weights, macro-avg AUC; [21] – converting Likert scales into bins (adjustment of ranges) |
| | balanced data | [24], [48], [49], [51], [56], [58], [59], [64], [46]?, [42]?, [50]?, [53]? |
| | imbalanced data | [21], [26], [31], [40], [42], [44], [45], [47], [57], [65], [23]? |
| Statistical tests on results | none | [19], [20], [24], [26], [31], [40]–[43], [45]–[53], [56], [57], [59], [61], [62], [64], [65] |
| | applied | [54], [55] – ANOVA, LSD; [63] – p-value, analysis of beta coefficient; [58] – p-value; [60] – McNemar's test (within-subjects chi-squared test); [23] – ANOVA; [44] – Wilcoxon signed-rank; [21] – Pearson correlation coefficient; [22] – student's t-tests |

discrete emotions. For example, 1={*sad*}, 0={*angry, fear, happy, relax*} [57].

Yet another binary approach was presented in [31], [51]: combinations of every discrete emotion against each other, i.e., one-versus-another. Interestingly, it referred both the initial discrete [31] and 1-dimensional model [51]. A separate solution was investigated by Schmidt et al. in [21]. They gathered self-assessments for more than one affective state, i.e., also for stress and anxiety. Based on that, they tested multi-task deep learning models. Such classifiers are able to reason about multiple outputs while learning from only one common input.

To sum up, both dimensional and discrete initial models were most frequently converted to quite simple discrete or even binary ML problems. This may come from relatively few data samples collected in the studies, as it made multi-label or multiclass classification with a greater number of classes hardly feasible.

Only three papers considered continuous values and regression problems [22], [63], and probably [24] (deduced by us), while regression was used to evaluate impact of affect on productivity in [65].

### 5.9 Windowing and Case Definition

Different approaches to studies require different definitions of a learning case, which is the basic unit of knowledge supplied to the machine learning model. In lab studies, one learning case may be defined as, e.g., a signal traced during the whole stimulus exposition [56], a part of it [41], or across several stimuli of the same type combined into one longer stimulus [52] (the authors did not mention splitting the signal; hence, we deduce they used the whole recording). During field studies, the signal collection remains uninterrupted for a long time, and thus it is necessary to extract only their fragments. In field studies, extracting proper learning cases may prove challenging, as the beginning of stimulation is unknown, and its end cannot be pointed with much certainty. In the emotion recognition area, consistent segments of the registered physiological signal are often called *windows*. Windowing may be used to create many learning cases from one acquired signal by dividing it into parts (windows) and labeling each of them with the same label [58], [59]. Such a method of increasing the number of samples may prove useful when training deep learning models, as overall, they require vast amounts of data to work properly. For example, in [21], even 240 cases were extracted from each labeled affective state. Yet another way is to treat the divided signal as a time series and to feed it to the neural network with recurrent layer [62].

Although only 41% of the included papers contained any information regarding windowing, they considered many different aspects regarding the process. In four papers [22], [51], [55]–[57], the authors exploited windows comprising the whole recorded signal. In papers considering only fractions of recordings, the window size varied from as little as one second [23] to 180s or even more [51], [64]. Wampfler et al. [45] used two separate window sizes to extract features from two signal sources. In six papers, the authors included information whether the sliding windows were overlapping (shifted by fewer samples than window size) [21], [23], [44], [58], [59], [63] or adjacent [41], [62]. In two papers, the researchers used different window sizes, depending on the type of signal. Wampfler et al. [45] utilized windows of ten seconds for physiological signals and considered the whole stimulus for data from a stylus.

### 5.10 Signal Preprocessing

The role of signal preprocessing is to remove interferences and artifacts from physiological data that are not related to emotional patterns and may have a negative impact on the results of affective computing. The primary sources of such distortions are both external and internal. They include external electromagnetic fields influencing electronic circuits inside wearable devices, electronic noise generated inside the circuits, body movements, sliding and changes in contact of sensors on the body surface, mixing of signals generated by different organs (e.g., the brain, heart or muscle electrical activities), temporary deactivation or saturation of sensors, etc. The methods used to remove artifacts or restore damaged fragments are usually based

TABLE 9
Methods used for signal preprocessing

| Preprocessing method | Used by |
|---|---|
| Filtration and smoothing (lowpass, bandpass, notch, median, moving average, aggregation, mean or drift removal) | [20], [22], [23], [26], [31], [41], [44], [45], [51]–[58], [60], [63] |
| Decomposition and removing undesirable components (WT, ICA) | [23], [47], [54], [55] |
| Normalization (to [0,1] or [0,100] range, in relation to the relaxation state) | [19], [20], [41], [48], [49], [52] |
| Winsorization (removing outliers and dubious or corrupted fragments, thresholding) and interpolation of data | [41], [44], [45], [55], [61], [62] |

on the known properties of the monitored signals (such as shown in Tabs. 5 and 6) or are data-driven ones. They can be divided into four main groups, Tab. 9. Initial antialiasing filtering in the input circuit is necessary (according to the Nyquist-Shannon sampling theorem) before converting the processed voltage into a digital signal. Further filtration (both analog or digital) and smoothing favor the desired frequency components and reduce others. Generally, high-pass filters remove slowly wandering components, low-pass filters reduce high-frequency noises, pass-band filters focus on specific frequency ranges, and the signal smoothing methods put low-frequency components ahead of others. Decompositions enable distinguishing between the desired and undesirable components. Among them, the wavelet transform (WT), implemented in the continuous (CWT), discrete (DWT), or fast (FWT) form, converts a nonstationary signal into the coefficient components of different scale (also related to frequency ranges) by finding its local correlations with a specially selected family of wavelets. Typically, the first component may represent measurement noise or an EMG artifact so that it can be removed afterward. Another example is independent component analysis (ICA) used to separate independent source signals that are mixed while being recorded by different sensors, provided that the number of recorded signals is not less than that of the sources. Normalization makes it possible to keep the energy of signals at a comparable level, especially when the data from different devices are combined, and to extract features with adequate values. Finally, winsorization reduces extreme values or interpolation replenishes damaged fragments, taking into account the statistical properties of some or most of the adjacent data.

### 5.11 Feature Extraction and Selection

Most of the analyzed papers follow the classical approach to machine learning, requiring the separate steps of handcrafted feature extraction and then their further selection (Fig. 2). The extracted features represent the specific properties of selected segments of a signal or data covered by a sliding window. The number of extracted features is usually smaller than that of raw data, so the dimensionality of a problem is reduced while maintaining relevant information. Overall, the features are primarily extracted within three domains of: (1) time, (2) frequency, and (3) time-frequency or time-scale (for non-stationary signals), using appropriate transformations if necessary, Tab. 10. Among them, the most commonly used are the Fourier (FT), and

TABLE 10
Methods used for feature extraction

| Feature extraction method | | Used by |
|---|---|---|
| Time domain | Signal morphology (amplitude, extrema, intervals, etc.) | [19], [21]–[23], [31], [40]–[45], [47]–[49], [51], [54]–[57], [60], [61], [63]–[65] |
| | Hjorth parameters | [23] |
| | Rate of specific events | [20], [21], [26], [40], [41], [43]–[45], [54], [55], [63] |
| | RMS | [21], [24], [26], [43]–[45], [51], [54], [55], [61] |
| Frequency domain | PSD | [21]–[23], [26], [41], [43]–[45], [56], [57], [61] |
| | Frequency spectra | [22], [41], [44], [45], [52], [56], [57], [64] |
| Time-scale domain | WT, tonic and phasic components | [31], [52], [64] |
| Statistical indices | Mean, median, SD, skewness, kurtosis, correlation, etc. | [19], [21]–[24], [26], [31], [40], [41], [43]–[45], [48], [49], [51], [54]–[59], [61], [63] |
| Nonlinear measures | Measures of chaos, Poincare plots, entropy | [23], [41], [44], [51], [52], [61] |

TABLE 11
Methods used for feature selection

| Method | Reduction factor | Used by |
|---|---|---|
| PCA | 5/13 in [57], 35/1000 in [31] | [31], [41], [57], [61] |
| Correlations | 21/84 | [61] |
| SFS+SVM | (14-18)/28 | [41] |
| IG | 23/28 | [41] |
| MI | (5-20)/53 | [51] |

wavelet (WT) transforms, along with the decomposition of EDA into the tonic and phase components. Regardless of the target domain, specific scalar metrics are additionally superimposed on such signals or transformations, yielding the final set of extracted features. They can be classified as morphological properties, dynamic properties given by the Hjorth parameters, energetic parameters: root mean square (RMS), power spectral density (PSD), or statistical indices: mean value, median, standard deviation (SD), etc. It is worth noting that despite their different origins, RMS and SD represent the same information according to the way they are computed. In addition, due to the origin of physiological signals from the nonlinear dynamical systems, specific nonlinear measures are also applied to characterize them, such as indices of chaos or complexity, Poincare plots or entropy.

The last stage, consisting in the selection of a subset of discriminatory features, results in a further reduction of the dimensionality of the problem. It takes into account the redundancy of the previously extracted features or their inability to distinguish between considered emotional states. Since testing of the informative properties of all possible subsets (combinations) of features is usually impossible within a reasonable time, suboptimal methods are typically used to this end. There are a few basic schemes for feature selection: transformation, filtering, wrapper and embedded methods. The first one applies a linear projection to another space, where new (e.g., orthogonal) features are arranged according to a specific rule, e.g., decreasing variance in

principal component analysis (PCA), and only the first few of them are further used. Filtering relies on specific measures, such as correlations, information gain (IG), or mutual information (MI), together with thresholding to select the best features. For wrapping, selection is combined with a classification method (such as the SVM) to check the effects of successive subsets of features on the classification effectiveness. Finally, some deep learning algorithms automatically produce a new feature representation, so they are called embedded ones. The above approaches are usually combined with schemes for adding the best or removing the worst features, called the sequential forward (SFS) or backward (SBS) selection. Carrying feature selection out increases the efficiency of classification algorithms in terms of computational complexity and final accuracy. Unfortunately, this stage is not mentioned in most of the reviewed papers. The only methods used include: PCA, correlations, SFS with SVM, IG and MI, Tab. 11.

### 5.12 Reasoning Models

The most popular approach among different machine learning tasks was multiclass classification applied in 74% of papers, Tab. 8. In eight papers, researchers considered binary classification problems, often simplifying the psychological models of emotions used in their studies, Sec. 5.8. The least popular was solving a regression task, which was adopted in only three papers [22], [24], [63]. Schmidt et al. [21] were the only ones solving a multi-task problem and utilizing a single ML model to classify four different problems simultaneously.

Overall, simple ML models were the most popular in the included papers. Among them, SVM (41% of papers), KNN (29%), DT (29%), and random forest (RF, 26%) were implemented most frequently. In two papers [40], [56], rules for decision trees were hand-crafted based on the analysis of statistical features. Kanjo et al. [61] utilized the Stacking algorithm, creating an ensemble of KNN, SVM, and RF models, with Naive Bayes (NB) model as a learner. Classical machine learning approaches have not lost their popularity over the years. In 2020, they were still more common than deep learning architectures. In total, only 21% of studies applied deep learning algorithms. In early deep learning approaches, multilayer perceptrons (MLP) were mostly used [48], [49]. Kanjo et al. [62] and Nakisa et al. [23] still used MLP in 2018, but both also experimented with more advanced architectures like CNN or LSTM networks. In 2019, Schmidt et al. [21] experimented with CNNs in the end-to-end setup. In 2020, Saxena et al. [24] used a simple neural network for regression task, and Tizzano et al. [59] used an LSTM model in the transfer learning approach.

There was more than one architecture used in 56% of papers, see Supp. Mat. Tab. 6. In 63% of them, only simple classifiers were compared. In the remaining 37%, researchers exploited both classic machine learning and neural networks for emotion recognition. Out of these, four papers compared classic and deep models using the same feature sets and tasks [23], [41], [48], [49]. Only four papers examined more than one neural network architecture [21], [23], [59], [62].

### 5.13 Validation

In principle, we can distinguish six general approaches to validation of emotion recognition, which differ depending on the set of cases used to split into train and test examples, Tab. 12:

1) **Non-specialized validation**, taken from the general methods used in machine learning. They operate on the whole set of learning cases. They do not respect subject's specificity or stimuli invoking emotional reactions neither. These methods range from (a) the simple one-time split [24], [54], [55], [61], [64]; (b) stratified one-time split preserving output class distribution [57]; to (c) classic k-fold cross-validation [19] , Fig. 7A, or its variation – leave-one-out, i.e., repeated keeping one case for testing while learning on the remaining data [48], [49], [51].
2) **Intra-subject validation** a.k.a. **user-dependent validation**. The set related to a given subject is split into train and test data. Then, we learn on some cases about the subjects and test on their remaining examples. This is performed for all subjects collaboratively. All intra-subject validation methods assume that we need some data about a given subject available to train the classifier. It does not measure the ability of the model to reason for new subjects. Here, we can have: (a) the simple one-time split method [47], [62]; (b) intra-subject k-fold cross-validation [50], [58]; (c) leave-one-case-out for a given subject [22]; or (d) leave in the test data some cases related to a given stimuli or assessment, e.g., leave-one-video-out (LOVO), Fig. 7B [44], k-fold cross-validation over stimuli (video) [44] or leave-target-questionnaires-out (LTQO) [21].
3) **Inter-subject validation** a.k.a. **user-independent validation** is specific for reasoning from human-related data. It leaves all data related to some subjects in the testing collection while training on all other subjects. It provides more insight into the model's capability to recognize emotions for unknown people, which is a more realistic and practical setup. Some researchers rely on a one-time split [56]. However, most mature studies apply leave-one-subject-out validation (LOSO), Fig. 7C [21]–[23], [31], [41], [44], [45], [50], [52], [58]. In the case of the greater number of participants, the data of more than one subject may be left for testing, e.g., of two subjects [42].
4) **Inter-intra-subject validation** is suitable if the generalized model is combined with the personalized one. Then, leave-one-subject-out (LOSO) is applied for generalization purposes along with the repeated random split on data from the left subject [59].
5) **Task-based cross-validation**. The model is trained on subjects doing one task. The testing is done on the same subjects but during their other activities, Fig. 7D, [45].
6) **Across time validation** learns on data collected in one period. The testing is performed on examples from the same subjects but in the later periods, Fig. 7e, [20], [26].

Usually, *a learning case* here is a single annotation (class label) with the corresponding physiological signals. However, these signals can be split into windows providing multiple learning cases with the same label, e.g., to enhance the training of deep learning models [21], see also Sec. 5.9.

Both the set used to split into training and testing par-

TABLE 12
Reasoning model validation; division into training and test sets. The number of '-' or '+' denotes the ability to estimate generalization level of the considered models. '?' means it was not explicitly provided by the authors but inferred by us

| Set for split | Validation type | Generalizability | Used by: Details |
|---|---|---|---|
| Whole set | Validation on the whole set | - - - - | [60] |
| | One-time split over all cases | - - - | 75% train, 25% test or 70% train, 15% validation and 15% test [24]; 70% train, 30% test [61], [64]; 50% train, 50% test [54], [55] |
| | Stratified one-time split over cases | - - | $^1/_3$ train, $^2/_3$ test [57] |
| | Repeated random split over all cases | - | 10 random splits 70% train, 30% test [59] |
| | Classical k-fold or leave-one-out (LOO) cross-validation over cases | - | 10-fold [19]; ?-fold [43]; LOO [48], [49], [51] |
| Intra-subject | One-time split over all cases of each subject | - - | 70% train, 30% test [47], [62] |
| | Intra-subject cross-validation | + | 10-fold, independently for each subject [50], [58]; leave-one-out (LOO) = leave-one-observation-out of a given subject [22] |
| | Leave-one(k)-assessment(stimuli)-out | + | Leave-One-Video-Out (LOVO), 10-fold cross validation over video [44]; leave-target-questionnaires-out (LTQO), a stratified N-fold split over classes and questionnaires (80%/10%/10% - train/test/validation) [21] |
| Tasks | Between the different tasks performed | ++ | math tasks - training, watching pictures - testing [45] |
| Time | Across time validation | ++ | 1st week - training, 2nd and 3rd weeks - testing [20]; first three weeks - training, 4th week - testing [26] |
| Inter-subject | One-time split over subjects | ++ | 21 subjects - train, 5 subjects - test [56] |
| | Leave-one($k$)-subject-out (LOSO) | +++ | LOSO [21]–[23], [44], [45], [50], [52], [58], [59], [41]?, [31]?; two-subjects-out [42] |
| Inter-intra-subject | LOSO + intra-subject repeated random split | +++ | [59] |
| No info | No info | N/A | [20], [40], [46], [53], [63], [65] |

tition, as well as the splitting method, directly impact on the ability to assess the generalizability of the investigated approach, Tab. 12. This, in turn, shows to what extent we can apply the solution to real life or in other scientific studies.

Some domain-specific cross-validation approaches, especially inter- but also intra-subject methods, are better than classical cross-validation. They respect a kind of context in emotion recognition [91], e.g., by learning on some stimuli and testing on the other ones: leave-one-video-out [44].

Only [59], [62] address a vital question whether to train a general model for all subjects or to build multiple personalized classifiers adjusted to each individual.

Some researchers considered but abandoned the inter-subject validation since they did not achieve satisfactory performance, *"due to the high inter-subject variability that affects the physiological signals"* [44].

Overall, inter-subject validation, especially leave-one(k)-subject-out (LOSO) a.k.a. leave-one-proband-out (LOPO) [22], appears to better reflect real-world inference, in which classification models should not overfit ("over-train") the training data collected from participants recruited to a given study. This subject-oriented validation procedure is thought of as the most suitable (SOTA) for research on emotion recognition. It was used in 35% papers (12 out of 34). Note they were all published quite recently, in 2018 and beyond, even though the LOSO concept for validation is not new – it was utilized already in 1993 [92].

A very interesting and promising is inter-intra-subject validation, which is appropriate for models that combine (1) components trained on data from all but one subject (inter-subject LOSO validation) with (2) layers learned on some data from the left subject (intra-subject validation). Then, the testing is performed on the remaining data of the left subject. The split within the remaining subject may be repeated at random [59] or alternatively done as k-fold cross-validation (not considered in any paper).

A properly performed validation procedure provides some information about the generalization ability of the considered methods worked out using the given data. It is measurable, e.g., with standard deviation value calculated over cross-validation folds. Moreover, in the case of inter-subject validation, e.g., LOSO, such values inform with what margin a tested solution (model) is able to predict emotions for a new subject. Unfortunately, only some researchers deliver any information about the generalization of their methods. For example, among 12 included papers that used leave-one($k$)-subject-out (LOSO) cross-validation, five did not report any information related to differences between subjects, e.g., standard deviation, individual values for each subject, mean absolute error, root mean square error [31], [41], [42], [52], [59]. Hence, their proper cross-validation was only partially exploited.

## 6 Discussion

### 6.1 Study Design

There was no research in SLR that could be seen as comprehensive and reliable with respect to all research components discussed in Sec. 5. In particular, (1) only five studies (15%) investigated more than four emotions (five/six discrete ones or six detected emotional regions); (2) the approval by an ethical committee or workers council was reported in only nine papers (26%); (3) the authors of only 14 papers (41%) collected written participants' consent; (4) only ten papers (29%) considered any context data, e.g., alcohol consumption, physical exercises, out of which two induced physical activity; (5) any subjects' health conditions, e.g., personality

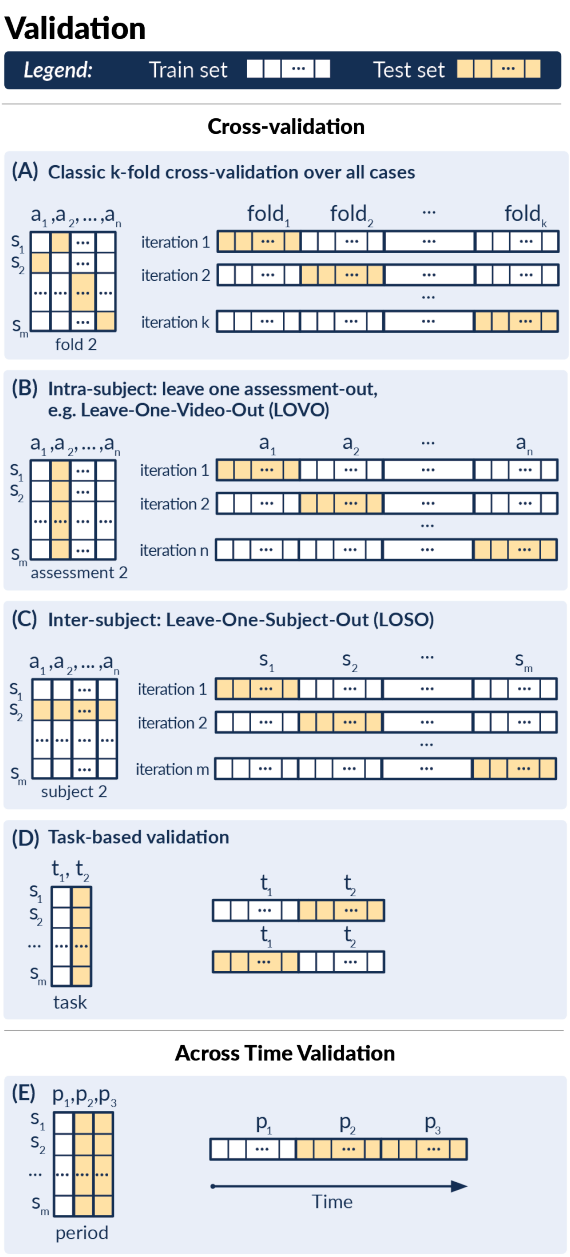


Fig. 7. Selected validation methods used in emotion recognition. $a_i$ denotes the *i*th assessment, $s_j$ – the *j*th subject. The matrix may not be fully filled out, since some examples may be removed or uncollected.

disorders or cardiovascular diseases, were checked in only 14 papers (41%); (6) in three papers (9%) labeling procedure cannot be verified as it was not described at all, while labels assigned to stimuli were not verified by self-assessment in seven papers (21%); (7) only three out of 11 papers having imbalanced data applied appropriate techniques to deal with this problem; (8) only seven studies (21%) performed any adjustment of reasoning model parameters, out of which only three applied parameter optimization; (9) inter-subject validation procedure, which can be treated as SOTA, was applied in only 12 papers (35%); (10) results in only nine papers were tested statistically (26%); (11) no research, which acquired their own emotionally labeled physiological signals, shared the data; (12) software source code was published together with only one paper. Besides, more than 50 subjects were investigated in four papers (12%). None of them exceeded 100 people. It means that most studies are rather small in scale.

Most of the SLR studies were carried out in the lab conditions using pre-defined, dedicated stimuli. However, the real-life environment explored in eight papers poses different challenges, Tab. 2. In particular, researchers have to solve the labeling problem in a convenient and non-annoying way, keeping in mind that some emotions like fear or disgust do not happen often, leading to high class imbalance.

### 6.2 Emotional Models and ML Problems

Simple emotional and affective models like *low arousal-high arousal* or *no stress-low stress-high stress* were extensively explored in the literature. It mainly resulted from the strong correlation between arousal or stress and some biological signals – GSR or BVP [4], [93]. Nevertheless, emotions are much more complex, and their multidimensional nature remains a great challenge for future work. Most SLR studies focused on only few basic emotions, up to six, Supp. Mat. Tab. 5. However, Du et al. [94] identified as many as 21 emotions from facial expressions, whereas Cowen and Keltner [66] found 27 distinct emotional categories from self-reports. A greater number of emotions to distinguish from each other makes the recognition problem much more difficult.

The complexity of emotions also appears to be an additional reason why in only a few studies, the authors monitor and recognize emotions in the non-controlled (field study) or semi-controlled environment (field with constraints).

The emotion model directly impacts the detection model. Multidimensional models should straightforwardly lead to a multi-label classification problem, which in turn requires much more cases to train the classifier. None of the papers approach reasoning in such a comprehensive way. Please note that correct multi-label models would be able to recognize combinations of emotions that even do not occur in the training set. Nevertheless, it was partially solved in seven papers utilizing dimensional emotional models and multiple independent regressors [22] or classifiers [21], usually many binary ones [19], [41], [43], [44], [51].

Building complex models that combine general and personal knowledge appears to be a future challenge [59]. It especially refers to deep learning architectures that need to separately learn on the whole population of subjects and on personal signals gathered from a given person. Moreover, the specificity of individual human physiology requires more personal rather than generalized solutions. It holds true also for validation procedures. Please note that non-inter-subject validation methods usually provide relatively high-quality results since the models are trained and tested on data gathered from the same subjects [59].

### 6.3 Data Collection

Gathering data and labeling is a great challenge, especially in the wild. Larger amounts of data are particularly important for deep learning models that require many samples for training. Therefore, personal, multipurpose devices, as well as more sophisticated triggering methods [38], [39], may be a good solution for field studies.

The low variety of hardware utilized in the included papers, see Supp. Mat. Tab. 7, might stem from the fact

that most off-the-shelf wearables do not provide access to raw physiological signals. Additionally, wearables usually provide low-frequency signal sampling, and their signal quality is worse compared to the medical-level devices, especially when the user is in motion [86].

We focused on portable wireless wearables that can potentially monitor physiological signals in the field. Notwithstanding, some other signals and data may be complementary, e.g., data gathered by our smartphone about our activity [95], voice [96], or smartwatch built-in camera monitoring our face [97]. It seems that such multimodal approaches may provide more accurate results, but it also raises issues of data synchronization across multiple sensors and/or devices. Moreover, other modalities may be difficult to collect in everyday life. It especially refers to facial expressions. Voice monitoring, however, may be more feasible, see Amazon Halo wearable [98]. Continuous monitoring of physiological signals poses a problem with high energy consumption and the necessity to recharge the device frequently [39]. This is also valid while voice overhearing [98].

### 6.4 Data Processing

The articles under review demonstrate the crucial role of data preprocessing when physiological signals are recorded with ubiquitous wearable sensors, as they are typically heavily corrupted by artifacts masking useful affective information. Thus, it is necessary to recover it using, e.g., filtration, smoothing, normalization, winsorization, interpolation, or their various combinations.

The feature extraction methods retrieved from the analyzed works only partially illustrate the spectrum of such approaches that can be found in the analogous but wider literature. The main difference is the fairly limited number of methods that can decompose a physiological signal into several specific components, including in addition to the used wavelet transform, also the empirical mode decomposition (EMD), Hilbert-Huang transform (HHT), variational mode decomposition (VMD), or matching pursuit algorithm (MPA).

The most surprising finding in the area of physiological data processing is the almost complete abandonment of feature selection, although it can reduce the computational complexity of machine learning algorithms and increase their accuracy. The methods worth considering in future studies would be also: relief algorithms, analysis of variance (ANOVA), or minimal redundancy maximal relevance procedure (MRMR).

The duration of emotions and thus the size of windows with signals considered remains unsolved. It should be studied more deeply, especially in the field research, for which we do not control stimuli.

### 6.5 Machine Learning

The majority of papers (76%) apply plain classifiers based on decision trees or hyperplanes (SVM, LDA, etc.), which seem to be efficient in simple binary problems. Several works, however, proved that more advanced deep learning architectures are better suited to model complex emotional data, in particular multimodal [21], [23], [59], [62]. We found that the deep learning approach, especially end-to-end [21], [62], [89], was explored very little, while it has a great potential to boost emotion recognition.

We are concerned that many papers did not consider imbalance in learning samples when it was apparent that such problem exists [23], [26], [31], [40], [42], [47]. Even worse, the authors utilized accuracy measure to assess the quality of classification. In such a case, the results were most probably overestimated.

Furthermore, the quality measure is commonly used to evaluate the model efficiency. We should keep in mind that such a measure is only a numerical value that represents how well we are able to model the relation between physiological signals and questionnaires filled out by the participants during the study. The quality measure cannot describe how well the model will actually perform in everyday life, where we have a full range of emotions, often co-occurring, e.g., fear and anger when we drop our smartphone on the floor. In such a case, the model will not be able to determine a class label (emotion) that has not been included in the training data, or, alternatively, it will wrongly suggest the most probable class it knows. Such mistakes will significantly reduce the quality measure level.

### 6.6 Research Replicability and Comparability

Based on our SLR, we found it is virtually impossible to compare different studies because:

1) they operate on different emotional models, i.e., various basic discrete emotions or emotional dimensions are not equivalent, e.g., both [47] and [40] utilized four discrete emotions, but does *fear* from [47] correspond to *pain* in [40]? (Sec. 5.1);
2) their emotional models are transformed to different ML problems (Sec. 5.8);
3) they use different kinds of data, i.e., various signals gathered from distinct sensors and wearables with different quality, including sampling frequency (Sec. 5.5, Supp. Mat. Sec. 5.6);
4) they apply incomparable quality measures and validation procedures (Sec. 5.13 and Supp. Mat. Sec. 3.7);
5) their datasets and software code are not available for other researchers (Supp. Mat. Sec. 4).

Having the above obstacles identified, we postulate to additionally apply the common (1) emotional models (e.g., Ekman-Friesen [68], Plutchik [70] or dimensional arousal-valence, as they are), (2) validation procedures (e.g., LOSO), and (3) quality measures (e.g., F-measure, ROC AUC, accuracy). It means that we encourage the researchers to refer to the established methods and models rather than only focus on their own solutions. We also call for the publication of data and source codes to allow replicability.

### 6.7 Challenges and Future Research Directions

The emotions experienced in real life have a complex nature, and the subjects pretty often perceive more than only one pre-defined basic emotion [24]. Therefore, especially in field studies, we can consider emotions as constructed on the fly by individuals depending on the recent context [99]. It can be achieved by gathering more free text self-assessments rather than fixed questionnaires, as well as more contextual metadata. Besides, *multi-label classification* or *multivariate regression* [24] should be applied to achieve outputs with multiple classes/values, i.e., multiple emotions simultaneously, as emotions experienced in real-life

usually co-occur. Additionally, assuming that emotions are only one of many affective states, thus, they are interlinked with stress, anxiety, flow, and mood, then the dedicated multi-task deep learning models are able to efficiently make use of their mutual relationships [21].

We believe that the recent breakthroughs in deep learning open huge possibilities for automatic signal processing and representation. For example, the autoencoders might be used to approximate missing data, thus, allowing for better signal quality. The generative adversarial networks may be applied to simulate a continuous race between sympathetic and parasympathetic nervous systems. The long short-term memory neural network could be extended with the attention mechanism to better target a specific emotion. Perhaps the artificial general intelligence (AGI) with its concept of whole-brain emulation [100] will allow for an even more complex representation of human processes.

Personalization of reasoning is definitely a great challenge for future development. It respects significant individual differences in both physiology and in the experience of emotions. Therefore, an additional interesting task is how to fuse the knowledge extracted from the entire population (general models) with the specificity of individual subjects (personal models) [59]. This, in turn, requires new inter-intra-subject validation methods that would test both the generalization and personalization abilities of the complex model. General and personalized solutions also demand appropriate active learning methods to update models according to new data acquired independently from an individual and all subjects.

Finally, the reasoning models trained on data collected in the laboratory environment will be inefficient in the uncontrolled everyday life scenario. Therefore, we should gather annotated training data in the field, where emotions appear in the natural context. For that purpose, we need precise and reliable sensors with an appropriately high sampling frequency and possibly with the ability to collect multimodal data, e.g., BVP+EDA or physiological signals and speech data.

While considering studies in the field, also differences between emotions experienced in various contexts like physical or sporting activities, e.g., hiking or eating should be taken into account. It also includes compensation of biases caused by the temporal context. Analysis of these factors should also be supported by experts in the field.

We find large-scale studies in the wild and open science (data, source codes) crucial for future research.

## 7 Conclusions

In our systematic literature review, we have investigated over 3,000 articles to address the question of whether wearables can be used to recognize emotions in everyday life. Based on the 34 relevant papers, we believe that such a solution is feasible but still requires further investigation.

We were able to identify seven scenarios for emotion recognition from physiology. It is observable that studies slowly transfer from the laboratory setup to the field environment. This demands, however, comfortable devices able to provide reliable measurements in motion and user-friendly self-assessment triggering procedures.

Our positive findings are: (1) emotion recognition for everyday life using physiological signals from wearables is possible; (2) there are many off-the-shelf wearables devices equipped for field investigations; (3) more and more studies apply inter-subject validation, e.g., leave-one-subject-out; (4) deep learning architectures provide new opportunities to solve complex tasks in field studies.

The negative conclusions of the SLR are: (1) the majority of studies were performed in the lab; (2) most of them simplified classification to the binary or few-class problem; (3) the machine learning stage was often neglected, in particular, imbalance of learning samples was ignored; (4) it is almost impossible to reproduce any research.

The SLR results and other resources useful in the emotion recognition task will continue to be updated in the Emognition portal [16].

## Acknowledgments

The authors would like to thank members of the Emognition research group, who participated in the initial review of the SLR articles: Magda Boruch, Anna Dutkowiak, Maciej Dzieżyc, Jakub Dziwiński, Katarzyna Jabłońska, Patrycja Jakimów, Joanna Komoszyńska, Dominika Kunc, Weronika Michalska; also to Beata Trawińska for her help with the figures. This work was partially supported by the National Science Centre, Poland, project no. 2020/37/B/ST6/03806; by the statutory funds of the Department of Artificial Intelligence, Wroclaw University of Science and Technology; by the Polish Ministry of Education and Science – the CLARIN-PL Project.

## References

[1] J. S. Lerner, Y. Li, P. Valdesolo, and K. S. Kassam, "Emotion and decision making," *Annu. Rev. Psychol.*, vol. 66, pp. 799–823, 2015.
[2] D. Kahneman, *Thinking, fast and slow*. Macmillan, 2011.
[3] C. A. Smith and R. S. Lazarus, "Emotion and adaptation," in *Handbook of personality: Theory and research*, 1990, pp. 609–637.
[4] P. Schmidt, A. Reiss, R. Dürichen, and K. V. Laerhoven, "Wearable-based affect recognition—a review," *Sensors*, vol. 19, no. 19, p. 4079, 2019.
[5] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *ICMI 2018 - Proc. 2018 Int. Conf. Multimodal Interact.*, 2018, pp. 400–408.
[6] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, 2020.
[7] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.
[8] E. P. Torres, E. A. Torres, M. Hernández-Álvarez, and S. G. Yoo, "Eeg-based bci emotion recognition: A survey," *Sensors*, vol. 20, no. 18, 2020.
[9] C. Marechal, D. Mikołajewski, K. Tyburek, P. Prokopowicz, L. Bougueroua, C. Ancourt, and K. Węgrzyn-Wolska, "Survey on ai-based multimodal methods for emotion detection," in *High-Perform. Model. Simul. Big Data Appl.* Springer, 2019, pp. 307–324.
[10] M. Meyer, P. Helmholz, M. Rupprecht, J. Seemann, T. Tönnishoff, and S. Robra-Bissantz, "From the inside out: A literature review on possibilities of mobile emotion measurement and recognition." in *Bled eConference*, 2019, p. 23.
[11] A. Chunawale and M. Bedekar, "Human emotion recognition using physiological signals: A survey," *2nd Int. Conf. Comm. Inf. Proc. (ICCIP)*, 2020.
[12] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, p. 592, 2020.
[13] E. Maria, L. Matthias, and H. Sten, "Emotion recognition from physiological signal analysis: A review," *Electronic Notes in Theoretical Computer Science*, vol. 343, pp. 35–55, 2019.

[14] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, 2018.

[15] S. Saganowski, A. Dutkowiak, A. Dziadek, M. Dzieżyc, J. Komoszyńska, W. Michalska, A. Polak, M. Ujma, and P. Kazienko, "Emotion recognition using wearables: A systematic literature review-work-in-progress," in *2020 IEEE Int. Conf. Pervasive Comput. Commun. Workshops*. IEEE, 2020, pp. 1–6.

[16] "Emognition portal," 2021, accessed: 2021-03-16. [Online]. Available: https://www.emognition.pwr.edu.pl

[17] R. W. Picard, *Affective computing*. MIT press, 2000.

[18] Y. S. Can, B. Arnrich, and C. Ersoy, "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey," *Journal of biomedical informatics*, vol. 92, p. 103139, 2019.

[19] F. Setiawan, S. A. Khowaja, A. G. Prabono, B. N. Yahya, and S.-L. Lee, "A framework for real time emotion recognition based on human ans using pervasive device," in *2018 IEEE 42nd Annu. Comp., Softw., and Appl. Conf.*, vol. 1. IEEE, 2018, pp. 805–806.

[20] M.-S. Dao, D.-T. Dang-Nguyen, A. Kasem, and H. Tran-The, "Healthyclassroom-a proof-of-concept study for discovering students' daily moods and classroom emotions to enhance a learning-teaching process using heterogeneous sensors," in *Proc. of the Int. Conf. on Pattern Recognit App. and Methods (ICPRAM)*. Scitepress–Science and Technology Publications, 2018, pp. –.

[21] P. Schmidt, R. Dürichen, A. Reiss, K. Van Laerhoven, and T. Plötz, "Multi-target affect detection in the wild: an exploratory study," in *Proc. 23rd Int. Symp. Wearable Comput.*, 2019, pp. 211–219.

[22] T. Martens, M. Niemann, and U. Dick, "Sensor measures of affective leaning," *Front. Psychol.*, vol. 11, 2020.

[23] B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire, and V. Chandran, "Long short term memory hyperparameter optimization for a neural network based emotion recognition framework," *IEEE Access*, vol. 6, pp. 49 325–49 338, 2018.

[24] P. Saxena, S. Dabas, D. Saxena, N. Ramachandran, and S. I. Ahamed, "Reconstructing compound affective states using physiological sensor data," in *2020 IEEE 44th Annu. Comp., Softw., and Appl. Conf.* IEEE, 2020, pp. 1241–1249.

[25] M. Maier, C. Marouane, and D. Elsner, "Deepflow: Detecting optimal user experience from physiological data using deep neural networks," in *Proc. Int. Jt. Conf. Auton. Agents Multiagent Syst. AAMAS*, 2019, pp. 2108–2110.

[26] A. Exler, A. Schankin, C. Klebsattel, and M. Beigl, "A wearable system for mood assessment considering smartphone features and data from mobile ecgs," in *Adjun. Proc. 2016 ACM Int. Joint Conf. Symp. Perv. Ubiq. Comp. Wear. Comp.*, 2016, pp. 1153–1161.

[27] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biological psychology*, vol. 84, no. 3, pp. 394–421, 2010.

[28] H. Schlosberg, "Three dimensions of emotion." *Psychological review*, vol. 61, no. 2, p. 81, 1954.

[29] P. Ekman, R. W. Levenson, and W. V. Friesen, "Autonomic nervous system activity distinguishes among emotions," *science*, vol. 221, no. 4616, pp. 1208–1210, 1983.

[30] R. W. Levenson, P. Ekman, and W. V. Friesen, "Voluntary facial action generates emotion-specific autonomic nervous system activity," *Psychophysiology*, vol. 27, no. 4, pp. 363–384, 1990.

[31] H. Feng, H. M. Golshan, and M. H. Mahoor, "A wavelet-based approach to emotion classification using eda signals," *Expert Systems with Applications*, vol. 112, pp. 77–86, 2018.

[32] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE transactions on affective computing*, 2018.

[33] J. Z. Lim, J. Mountstephens, and J. Teo, "Emotion recognition using eye-tracking: taxonomy, review and current challenges," *Sensors*, vol. 20, no. 8, p. 2384, 2020.

[34] B. Kitchenham, "Procedures for undertaking systematic reviews: Joint technical report," *Comput. Sci. Department, Keele University (TR/SE-0401) and National ICT Australia Ltd.(0400011T. 1)*, 2004.

[35] P. Desmet, "Designing emotions," 2002.

[36] J. Hayano, T. Tanabiki, S. Iwata, K. Abe, and E. Yuda, "Estimation of emotions by wearable biometric sensors under daily activities," in *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*. IEEE, 2018, pp. 240–241.

[37] A. Fortin-Côté, N. Beaudin-Gagnon, A. Campeau-Lecours, S. Tremblay, and P. L. Jackson, "Affective computing out-of-the-lab: The cost of low cost," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 4137–4142.

[38] S. Saganowski, M. Behnke, J. Komoszyńska, D. Kunc, B. Perz, and P. Kazienko, "A system for collecting emotionally annotated physiological signals in daily life using wearables," in *Int. Conf. Affect. Comput. Intell. Interact. (ACII 2021)*. IEEE, 2021, pp. 1–3.

[39] M. Dzieżyc, J. Komoszyńska, S. Saganowski, M. Boruch, J. Dziwiński, K. Jabłońska, D. Kunc, and P. Kazienko, "How to catch them all? enhanced data collection for emotion recognition in the field," in *2021 IEEE Int. Conf. Pervasive Comput. Commun. Workshops*. IEEE, 2021, pp. 348–351.

[40] D. Pollreisz and N. TaheriNejad, "A simple algorithm for emotion recognition, using physiological signals of a smart watch," in *Proc IEEE Eng Med Biol Soc (EMBC)*. IEEE, 2017, pp. 2353–2356.

[41] B. Zhao, Z. Wang, Z. Yu, and B. Guo, "Emotionsense: Emotion recognition based on wearable wristband," in *SmartWorld, Ubiq. Intell., Internet of People and Smart City Innovation*. IEEE, 2018, pp. 346–355.

[42] A. Albraikan, D. P. Tobón, and A. El Saddik, "Toward user-independent emotion recognition using physiological signals," *IEEE Sensors Journal*, vol. 19, no. 19, pp. 8402–8412, 2018.

[43] M. Ragot, N. Martin, S. Em, N. Pallamin, and J.-M. Diverrez, "Emotion recognition using physiological signals: laboratory vs. wearable sensors," in *Advances in Human Factors in Wearable Technologies and Game Design*. Springer, 2018, pp. 15–22.

[44] L. Romeo, A. Cavallo, L. Pepa, N. Berthouze, and M. Pontil, "Multiple instance learning for emotion recognition using physiological signals," *IEEE Transactions on Affective Computing*, 2019.

[45] R. Wampfler, S. Klingler, B. Solenthaler, V. Schinazi, and M. Gross, "Affective state prediction in a mobile setting using wearable biometric sensors and stylus," in *Proceedings of The 12th International Conf. on Educational Data Mining (EDM 2019)*, 2019, pp. 198–207.

[46] F. Nasoz, O. Ozyer, C. L. Lisetti, and N. Finkelstein, "Multimodal affective driver interfaces for future cars," in *Proc ACM Int Conf Multimed*. ACM, 2002, pp. 319–322.

[47] L. Hu, J. Yang, M. Chen, Y. Qian, and J. J. Rodrigues, "Scai-svsc: Smart clothing for effective interaction with a sustainable vital sign collection," *Fut Gen Comp Sys*, vol. 86, pp. 329–338, 2018.

[48] C. L. Lisetti and F. Nasoz, "Using noninvasive wearable computers to recognize human emotions from physiological signals," *EURASIP J. Adv. Signal Process.*, vol. 2004, pp. 1672–1687, 2004.

[49] ——, "Categorizing autonomic nervous system (ans) emotional signals using bio-sensors for hri within the maui paradigm," in *ROMAN 2006 - IEEE Int. Symp. Robot Hum. Interact. Commun.* IEEE, 2006, pp. 277–284.

[50] A. F. Bulagang, J. Mountstephens, and J. T. T. Wi, "Tuning support vector machines for improving four-class emotion classification in virtual reality (vr) using heart rate features," in *J. Phys. Conf. Ser.*, vol. 1529. IOP Publishing, 2020, p. 052069.

[51] L. Shu, Y. Yu, W. Chen, H. Hua, Q. Li, J. Jin, and X. Xu, "Wearable emotion recognition using heart rate data from a smart bracelet," *Sensors*, vol. 20, no. 3, p. 718, 2020.

[52] F. Setiawan, A. G. Prabono, S. A. Khowaja, W. Kim, K. Park, B. N. Yahya, S.-L. Lee, and J. P. Hong, "Fine-grained emotion recognition: fusion of physiological signals and facial expressions on spontaneous emotion corpus," *International Journal of Ad Hoc and Ubiquitous Computing*, vol. 35, no. 3, pp. 162–178, 2020.

[53] A. J. Majumder, J. W. Dedmondt, S. Jones, and A. A. Asif, "A smart cyber-human system to support mental well-being through social engagement," in *2020 IEEE 44th Annu. Comp., Softw., and Appl. Conf.* IEEE, 2020, pp. 1050–1058.

[54] K. Rattanyu, M. Ohkura, and M. Mizukawa, "Emotion monitoring from physiological signals for service robots in the living space," in *ICCAS 2010*. IEEE, 2010, pp. 580–583.

[55] K. Rattanyu and M. Mizukawa, "Emotion recognition using biological signal in intelligent space," in *International Conference on Human-Computer Interaction*. Springer, 2011, pp. 586–592.

[56] H. W. Guo, Y. S. Huang, J. C. Chien, and J. S. Shieh, "Short-term analysis of heart rate variability for emotion recognition via a wearable ecg device," in *2015 Int. Conf. Intell. Inform. Biomed. Sci. (ICIIBMS)*. IEEE, 2015, pp. 262–265.

[57] H.-W. Guo, Y.-S. Huang, C.-H. Lin, J.-C. Chien, K. Haraikawa, and J.-S. Shieh, "Heart rate variability signal features for emotion recognition by using principal component analysis and support vectors machine," in *2016 IEEE 16th Int. Conf. Bioinform. Bioeng. (BIBE)*. IEEE, 2016, pp. 274–277.

[58] J. C. Quiroz, E. Geangu, and M. H. Yong, "Emotion recognition using smart watch sensor data: Mixed-design study," *JMIR mental health*, vol. 5, no. 3, p. e10153, 2018.

[59] G. R. Tizzano, M. Spezialetti, and S. Rossi, "A deep learning approach for mood recognition from wearable data," in *2020 IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*. IEEE, 2020, pp. 1–5.

[60] A. Albraikan, B. Hafidh, and A. El Saddik, "iaware: A real-time emotional biofeedback system based on physiological signals," *IEEE Access*, vol. 6, pp. 78 780–78 789, 2018.

[61] E. Kanjo, E. M. Younis, and N. Sherkat, "Towards unravelling the relationship between on-body, environmental and emotion data using sensor information fusion approach," *Information Fusion*, vol. 40, pp. 18–31, 2018.

[62] E. Kanjo, E. M. Younis, and C. S. Ang, "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection," *Information Fusion*, vol. 49, pp. 46–56, 2019.

[63] S. Kim, K. Patra, A. Kim, K.-P. Lee, A. Segev, and U. Lee, "Sensors know which photos are memorable," in *Proc. 2017 CHI Conf. Ext. Abstr. Hum. Factors in Comput. Syst.*, 2017, pp. 2706–2713.

[64] N. T. Nguyen, N. V. Nguyen, M. H. T. Tran, and B. T. Nguyen, "A potential approach for emotion prediction using heart rate signals," in *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE, 2017, pp. 221–226.

[65] Y. Kadoya, M. S. R. Khan, S. Watanapongvanich, and P. Binnagan, "Emotional status and productivity: Evidence from the special economic zone in laos," *Sustainability*, vol. 12, no. 4, p. 1544, 2020.

[66] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proc. Acad. Nat. Sci.*, vol. 114, no. 38, pp. E7900–E7909, 2017.

[67] D. Krech, R. S. Crutchfield, and N. Livson, *Elements of psychology*, 3rd ed. New York: Alfred A. Knopf, 1974.

[68] P. Ekman and W. V. Friesen, *Facial action coding system: Investigator's guide*. Consulting Psychologists Press, 1978.

[69] C. Darwin, *The Expression of the Emotions in Man and Animals*. London: John Murray, 1872.

[70] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of emotion*. Elsevier, 1980, pp. 3–33.

[71] J. A. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[72] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *J Behav Ther Exp Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[73] C. Izard, "Basic emotions, natural kinds, emotion schemas, and a new paradigm," *Perspec Psych Sci*, vol. 2, no. 3, pp. 260–280, 2007.

[74] A. Ortony and T. J. Turner, "What's basic about basic emotions?" *Psychological review*, vol. 97, no. 3, p. 315, 1990.

[75] K. Oatley and P. N. Johnson-Laird, "Cognitive approaches to emotions," *Trends Cogn. Sci.*, vol. 18, no. 3, pp. 134–140, 2014.

[76] J. D. Morris, "Observations: Sam: The self-assessment manikin an efficient cross-cultural measurement of emotional response 1," *Journal of advertising research*, vol. 35, no. 6, pp. 63–68, 1995.

[77] A. Sonderegger, K. Heyden, A. Chavaillaz, and J. Sauer, "Anisam & aniavatar: Animated visualizations of affective states," in *Proc. 2016 CHI Conf. Hum. Factors Comput. Syst.*, 2016, pp. 4828–4837.

[78] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales." *J Pers Soc Psychol*, vol. 54, no. 6, p. 1063, 1988.

[79] P. Wilhelm and D. Schoebi, "Assessing mood in daily life: Structural validity, sensitivity to change, and reliability of a short-scale to measure three basic dimensions of mood," *European Journal of Psychological Assessment*, vol. 23, no. 4, pp. 258–267, 2007.

[80] J. M. Smyth and A. A. Stone, "Ecological momentary assessment research in behavioral medicine," *Journal of Happiness studies*, vol. 4, no. 1, pp. 35–52, 2003.

[81] P. Schmidt, A. Reiss, R. Dürichen, and K. Van Laerhoven, "Labelling affective states" in the wild" practical guidelines and lessons learned," in *Proc. 2018 ACM Int. Joint Conf. Symp. Perv. Ubiq. Comp. Wear. Comp.*, 2018, pp. 654–659.

[82] S. for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures, W. Boucsein, D. C. Fowles, S. Grimnes, G. Ben-Shakhar, W. T. Roth, M. E. Dawson, and D. L. Filion, "Publication recommendations for electrodermal measurements," *Psychophysiology*, vol. 49, no. 8, pp. 1017–1034, 2012.

[83] D. R. Bach, G. Flandin, K. J. Friston, and R. J. Dolan, "Modelling event-related skin conductance responses," *International Journal of Psychophysiology*, vol. 75, no. 3, pp. 349–356, 2010.

[84] L. Ciechanowski, A. Przegalinska, M. Magnuski, and P. Gloor, "In the shades of the uncanny valley: An experimental study of human–chatbot interaction," *Future Generation Computer Systems*, vol. 92, pp. 539–548, 2019.

[85] S. Saganowski, P. Kazienko, M. Dzieżyc, P. Jakimów, J. Komoszyńska, W. Michalska, A. Dutkowiak, A. Polak, A. Dziadek, and M. Ujma, "Consumer wearables and affective computing for wellbeing support," in *Proc. .17th EAI Int. Conf. Mob. Ubiquitous Syst.: Comput., Netw. Serv.* ACM, 2020, pp. –.

[86] A. A. Schuurmans, P. de Looff, K. S. Nijhof, C. Rosada, R. H. Scholte, A. Popma, and R. Otten, "Validity of the empatica e4 wristband to measure heart rate variability (hrv) parameters: a comparison to electrocardiography (ecg)," *Journal of medical systems*, vol. 44, no. 11, pp. 1–11, 2020.

[87] A. Borrego, J. Latorre, M. Alcañiz, and R. Llorens, "Reliability of the empatica e4 wristband to measure electrodermal activity to emotional stimuli," in *2019 International Conference on Virtual Rehabilitation (ICVR)*. IEEE, 2019, pp. 1–2.

[88] P. Konstantinou, A. Trigeorgi, C. Georgiou, A. T. Gloster, G. Panayiotou, and M. Karekla, "Comparing apples and oranges or different types of citrus fruits? using wearable versus stationary devices to analyze psychophysiological data," *Psychophysiology*, vol. 57, no. 5, p. e13551, 2020.

[89] M. Dzieżyc, M. Gjoreski, P. Kazienko, S. Saganowski, and M. Gams, "Can we ditch feature engineering? end-to-end deep learning for affect recognition from physiological sensor data," *Sensors*, vol. 20, no. 22, p. 6535, 2020.

[90] G. Keren, T. Kirschstein, E. Marchi, F. Ringeval, and B. Schuller, "End-to-end learning for dimensional emotion recognition from physiological signals," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 985–990.

[91] L. F. Barrett, B. Mesquita, and M. Gendron, "Context in emotion perception," *Current Directions in Psychological Science*, vol. 20, no. 5, pp. 286–290, 2011.

[92] M. S. Barlett, G. C. Littlewort, M. G. Frank, C. Lainscse, I. R. Fasel, and J. R. Movellan, "Automatic recognition of spontaneous facial actions," *American Psychologist*, vol. 48, pp. 384–392, 1993.

[93] J. Choi, B. Ahmed, and R. Gutierrez-Osuna, "Development and evaluation of an ambulatory stress monitor based on wearable sensors," *IEEE transactions on information technology in biomedicine*, vol. 16, no. 2, pp. 279–286, 2011.

[94] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[95] W. Sasaki, J. Nakazawa, and T. Okoshi, "Comparing esm timings for emotional estimation model with fine temporal granularity," in *Symp. Perv. Ubiq. Comp. Wear. Comp.* ACM, 2018, pp. 722–725.

[96] P. Denman, E. Lewis, S. Prasad, J. Healey, H. Syed, and L. Nachman, "Affsens: a mobile platform for capturing affect in context," in *Proc. of the 20th Int. Conf. on Human-Computer Interaction with Mobile Devices and Services Adjunct*. ACM, 2018, pp. 321–326.

[97] J. A. Rincon, A. Costa, P. Novais, V. Julian, and C. Carrascosa, "Intelligent wristbands for the automatic detection of emotional states for the elderly," in *Int. Conf. Intell. Data Eng. Autom. Lear.* Springer, 2018, pp. 520–530.

[98] K. Kozuch, "Amazon halo review," accessed: 2021-03-16. [Online]. Available: tomsguide.com/reviews/amazon-halo

[99] L. F. Barrett, *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, 2017.

[100] R. A. Koene, "Fundamentals of whole brain emulation: State, transition and update representations," *International Journal of Machine Consciousness*, vol. 4, no. 01, pp. 5–21, 2012.

**Stanisław Saganowski** is currently an Assistant Professor at Wroclaw University of Science and Technology, Poland. His research interests include emotion recognition, affective computing, and wearable sensors. In 2021, he was awarded a scholarship for outstanding young scientists by the Polish Ministry of Education and Science. He is a member of the Top 500 Innovators association founded by the Polish Ministry of Education and Science.

**Bartosz Perz** received M.Sc. degree in Big Data Analytics from Wrocław University of Science and Technology (WUST), Poland, in 2020. He is a Ph.D. student at WUST doing his studies with the Emognition team. His research interests include emotion recognition, affective computing and machine learning, and his main focus is on the personalized emotion recognition.

**Adam G. Polak** is an Associate Professor at Wroclaw University of Science and Technology and the Vice-Dean for Cooperation at the Faculty of Electronics, Photonics and Microsystems, WUST, Poland. His research includes: indirect measurements, modeling and measurements in physiology, biomedical signal processing, machine learning in disease detection. In 1992/1993 he held a fellowship from the British Council at the City University, London. He was a recipient of the awards of the Foundation for Polish Science (1995), Department IV of the Polish Academy of Sciences (PAS) (2008), and the scholarship “Enhancing science linkages between New Zealand and Europe through Poland” (2010). He is a member of the Committee on Metrology and Scientific Instrumentation, PAS, and an IEEE Member. Since 2018 he has been serving as an Associate Editor for the IEEE Transactions on Instrumentation and Measurement.

**Przemysław Kazienko** is a full professor of computer science and leader of ENGINE - the European Centre for Data Science and Emognition research team at Wroclaw University of Science and Technology, Poland. He has authored 300+ research papers, including 50 in journals with IF, related to affective computing and emotion recognition, sentiment analysis, hate speech, personlized NLP, social network analysis, spread of influence, and various machine learning problems. He initialized and led over 50 research projects with total budget 8M+ EUR. He gave 20 keynote/invited talks for international audience and served as a co-chair of 20+ international scientific conferences and workshops. He is an IEEE Senior Member, a member of the Editorial Board of several scientific journals, and also on the board of Network Science Society.