Computers in Biology and Medicine 166 (2023) 107489

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Wearable-based human flow experience recognition enhanced by transfer learning methods using emotion data

Muhammad Tausif Irshad [a,*], Frédéric Li [a], Muhammad Adeel Nisar [b], Xinyu Huang [a], Martje Buss [c], Leonie Kloep [c], Corinna Peifer [c], Barbara Kozusznik [d], Anita Pollak [d], Adrian Pyszka [e], Olaf Flak [f], Marcin Grzegorzek [a,g]

[a] Institute of Medical Informatics, University of Lübeck, Germany
[b] Department of IT, University of the Punjab, Lahore, Pakistan
[c] Department of Psychology, University of Lübeck, Germany
[d] Department of Social Science, Institute of Psychology, University of Silesia in Katowice, Poland
[e] Department of Human Resource Management, College of Management, University of Economics in Katowice, Poland
[f] Department of Management, Jan Kochanowski University of Kielce, Poland
[g] Department of Knowledge Engineering, University of Economics in Katowice, Poland

ARTICLE INFO

Keywords:
Flow
Human flow experience
Wearable sensors
Multimodal sensing
Physiological responses
Machine learning
Deep learning
Transfer learning
Artificial neural network

ABSTRACT

Background: Flow experience is a specific positive and affective state that occurs when humans are completely absorbed in an activity and forget everything else. This state can lead to high performance, well-being, and productivity at work. Few studies have been conducted to determine the human flow experience using physiological wearable sensor devices. Other studies rely on self-reported data.

Methods: In this article, we use physiological data collected from 25 subjects with multimodal sensing devices, in particular the *Empatica E4* wristband, the *Emotiv Epoc X* electroencephalography (EEG) headset, and the *Biosignalplux RespiBAN* – in arithmetic and reading tasks to automatically discriminate between flow and non-flow states using feature engineering and deep feature learning approaches. The most meaningful wearable device for flow detection is determined by comparing the performances of each device. We also investigate the connection between emotions and flow by testing transfer learning techniques involving an emotion recognition-related task on the source domain.

Results: The EEG sensor modalities yielded the best performances with an accuracy of 64.97%, and a macro *Averaged F1* (AF1) score of 64.95%. An accuracy of 73.63% and an AF1 score of 72.70% were obtained after fusing all sensor modalities from all devices. Additionally, our proposed transfer learning approach using emotional arousal classification on the DEAP dataset led to an increase in performances with an accuracy of 75.10% and an AF1 score of 74.92%.

Conclusion: The results of this study suggest that effective discrimination between flow and non-flow states is possible with multimodal sensor data. The success of transfer learning using the DEAP emotion dataset as a source domain indicates that emotions and flow are connected, and emotion recognition can be used as a latent task to enhance the performance of flow recognition.

1. Introduction

1.1. Background

Flow experience is a specific positive and affective state of mind that occurs when a person is completely immersed in an activity [1]. It is characterized as a state of self-forgetting and absorption during a task with demands that seem to match one's own skills exactly [2]. In addition, enjoyment is described as a key component of flow [3], as flow is usually perceived as a positive and rewarding experience [2]. In the past, psychologists have highlighted the impact of flow on

* Corresponding author.
E-mail addresses: m.irshad@uni-luebeck.de (M.T. Irshad), fr.li@uni-luebeck.de (F. Li), adeel.nisar@pucit.edu.pk (M.A. Nisar), x.huang@uni-luebeck.de (X. Huang), martje.buss@gmail.com (M. Buss), l.kloep@uni-luebeck.de (L. Kloep), corinna.peifer@uni-luebeck.de (C. Peifer), barbara.kozusznik@us.edu.pl (B. Kozusznik), anita.pollak@us.edu.pl (A. Pollak), adrian.pyszka@uekat.pl (A. Pyszka), olaf.flak@ujk.edu.pl (O. Flak), marcin.grzegorzek@uni-luebeck.de (M. Grzegorzek).

https://doi.org/10.1016/j.compbiomed.2023.107489
Received 26 May 2023; Received in revised form 9 August 2023; Accepted 15 September 2023
Available online 22 September 2023

psychological well-being [4,5], and flow has been shown to lead to positive outcomes in the workplace, such as high task performance and increased job satisfaction [6], which can bring benefits for both employees and employers [7]. The positive consequences of flow underline the relevance of measuring and recognizing it. However, the current standard for measuring flow consists of retrospective self-report questionnaires, which require interrupting participants in their tasks. This leads to an interruption of the flow experience and can restrict the potential positive consequences [8]. Therefore, psychological research has an ongoing debate about new methods to measure flow experience unobtrusively [8]. Measuring flow in an interruption-free manner would allow us to better understand the experience without interfering with the mechanisms themselves. In this way, flow situations can be better understood, anticipated, and actively created in order to benefit from the positive consequences of the experience. For instance, some questionnaire studies have already identified work tasks in which flow is experienced, such as planning, problem-solving, or evaluation tasks [9]. By measuring flow more precisely, flow situations could be better distinguished from non-flow situations. Then work tasks could be distributed and planned in a way that helps to increase the flow.

Furthermore, advances in wearable sensor technology combined with supervised machine learning (ML) models allow researchers to align self-reported data with objectively measurable physiological signals to develop algorithms that automatically recognize emotional and motivational states. Therefore, the automatic recognition of human flow experience in real-time from body-worn physiological sensor signals using supervised ML is a worthwhile investigation since ML techniques [10] have already shown reliable performances on such sensor signals in a variety of applications in different fields such as psychology, medicine, and biology [11–16].

A crucial step in ML is *feature extraction*, in which certain values relevant to the problem to be solved (i.e., *features*) are calculated from the sensor modalities data generally collected from the participants. Feature extraction methods transform the raw data into representative feature embeddings necessary for modeling. There are two main types of feature extraction methods – *feature engineering* and *deep feature learning* [12]. Feature engineering involves creating features manually, typically using specialized knowledge or basic transformation functions such as arithmetic operators and aggregation operators on sensor signals.

Deep feature learning approaches are based on *Artificial Neural Networks* (ANNs). ANNs belong to a class of mathematical models based on the concept of biological neural networks in the human brain. The main components of ANNs are artificial neurons – simple non-linear computational units connected in layers. A *Deep Neural Network* (DNN) is a derivative of an ANN, which contains many layers, including an input layer, an output layer, and various intermediate layers called hidden layers. In practice, DNNs are usually trained to map given input data to an output in a classification or regression task, depending on the problem to be solved [13,17]. They have become prevalent in recent years due to their ability to learn highly relevant features of a given problem [18].

1.2. Current challenges

The central challenge flow researchers are confronted with is that measuring flow with questionnaires interrupts the experience itself, and this interruption can lead to flow being gone for some time afterward [8]. In the worst case, flow research, as it is conducted in most cases, can result in less flow. To assess flow interruption-free and in real-time, new approaches to recognizing it by physiological sensor data have to be applied and investigated [19].

A natural attempt at analyzing flow in an interruption-free manner is to use unobtrusive wearable sensors collecting physiological data and then apply ML techniques on them to train flow recognition models. But in general, wearable-based flow experience recognition is a complex task because there is no state-of-the-art method to obtain accurate class label information from experts, such as in the video or image analysis domain. In contrast, in flow research, participants typically provide class label information during or after the end of each task via a questionnaire [8]. In addition, to our best knowledge, there is no public dataset available, and there are very few studies on this topic. The availability of a public dataset provides golden opportunities for scientists and researchers [20]. For example, it helps to reproduce the research results, enables scientists to build on others' work, and provides an opportunity to investigate the problem further, which can help improve accuracy or performance [21].

Most of the past flow literature studies focused on the game users' data to recognize flow when users played a video game, while a limited set of studies investigated flow in the working context. However, the literature suggests that playing games is a voluntary task that belongs to leisure or fun activities [22], and it is more probable for participants (or game users) to feel a flow state while playing compared to work activities that are not necessarily voluntary. Therefore, it is important to consider work-related activities and determine how flow is experienced at work.

From the machine learning point of view, only two studies so far have implemented deep feature learning approaches [23,24]. However, they also have some fundamental limitations, such as not implementing the cross-validation strategy in a way that the data of test participants could not be used for the training purpose and vice versa, and their presented results are mediocre (≈70% accuracy), which could be improved. Moreover, studies have yet to explore – how to circumvent the data scarcity problem in the flow recognition domain and how to use emotional data to enhance flow recognition performance.

1.3. Research motivation

The current challenges mentioned in Section 1.2 inspired the following motivations for this research. First, we wanted to investigate the question of interruption-free and objective flow experience recognition in the working context. As mentioned in Section 1.1, measuring flow in an interruption-free manner using unobtrusive wearable sensors would allow us to understand the experience better and, in this way, flow situations can be better anticipated and actively created in order to benefit from the positive consequences of the experience. However, there is currently no in-depth ML study in the related literature that aims to objectively recognize flow experience during working activities, and no public dataset related to this topic is available. This motivated us to fill this gap by collecting multimodal sensor data and performing ML experiments in a subject-independent manner.

ML has gained a lot of popularity in the past years due to the increasing availability of data and the impressive performances of ML models in some application domains, sometimes beating human performance. Deep learning has in particular become prevalent since deep feature learning approaches have shown to outperform feature engineering methods in several application fields, particularly those related to image processing [25]. However, ML approaches – more specifically, deep learning ones – are very reliant on large quantities of data to train robust models, which has caused progress in application fields relying on time-series data to be slow due to data scarcity problems. This is the case in particular for sensor-based flow experience recognition. To circumvent these issues, a subset of ML called *transfer learning* has received a lot of attention in the past decade. Transfer learning techniques attempt to extract knowledge from solving one task, referred to as the *source task*, and use it to improve the performance of a different but related task, referred to as the *target task* [26]. The idea behind such a transfer of information is that scarcity of data for the target task could potentially be mitigated by data available in large enough quantities for the source task. For this reason, transfer learning has become standard in applications using images due to the existence of powerful models trained using very large-scale datasets

such as *ImageNet* as the source dataset. But it remains significantly less explored for applications involving time-series data due to finding a suitable source dataset not being as obvious.

Our analysis of the flow literature revealed that past research had associated some emotion states with flow [2,27–30]. For instance, Csikszentmihalyi [2], the pioneer of flow theory, described flow as an optimal experience that can arise from a balance between the demands of an activity and one's own skills. If the demands are too high, a state of anxiety occurs instead; if, on the other hand, one's own skills exceed the demands, boredom arises [2]. This suggests moderate physiological arousal should promote flow, whereas low or excessive physiological arousal should impede it [28,30]. This association between flow and moderate arousal has already been observed in an experiment exposing participants to stressors and applying physiological and questionnaire measurements. An underlying u-shaped relationship between physiological arousal and flow is assumed, in which flow occurs when arousal is neither too low nor too high [27]. The aforementioned associations that researchers have found so far, combined with the availability of relatively large benchmark emotion datasets, such as the DEAP dataset [31], motivated us to conduct transfer learning experiments and try to leverage emotion data to improve the flow recognition performances that currently remain subpar, as mentioned in Section 1.2.

1.4. Main contributions

In this study, we hypothesize that modern multimodal wearable sensors with advanced ML models allow us to distinguish between human flow and non-flow states during work activities – and with higher accuracy than reported in the literature, particularly when enhanced by transfer learning using emotion recognition as a source task. Thus, we perform experiments to provide a comparative analysis of different state-of-the-art approaches to feature extraction and classification, either without any transfer or with a transfer of emotion-related knowledge to the problem of flow recognition.

The main contributions of this research are summarized as follows:

(1) Investigate the use of multimodal sensors in the context of human flow experience recognition and develop a state-of-the-art ML model, which learns flow and non-flow patterns from physiological responses and classifies them into their respective classes.
(2) Analyze and compare multimodal wearable devices' data and investigate how to circumvent the data scarcity problem in the flow recognition domain.
(3) Objectively investigate feature extraction approaches and ML algorithms for flow recognition by performing a comparative analysis between them in a subject-independent manner.
(4) Propose a deep transfer learning model for human flow state recognition using emotion classification as a source task and investigate which emotion-based source task helps to achieve enhanced flow recognition results.

The rest of the article is structured as follows. Section 2 presents the current state-of-the-art in human flow experience recognition. Section 3 describes the materials and methods used to analyze multimodal signals for assessing flow and non-flow states. Section 4 presents the experimental results. Section 5 provides a detailed discussion of our experimental results and findings. Finally, Section 6 concludes this work and provides suggestions for future studies.

2. Related works

Developing an unbiased system for predicting human flow experience when working with wearable multimodal sensor signals is a difficult task because, like emotions, flow is a subjective state that is not easy to decode [32]. In emotion analysis and flow detection, data are usually labeled by participants after each experiment, which may not be reliable and result in noisy labeling [33,34]. For subjective experiences like flow, there is – in particular, no possibility to acquire labels using unbiased external observers. Thus, there are few studies on this topic, and no public dataset is available. Most studies collected data when users played games, while others examined work-related activities. For example, Berta et al. [11] used a commercially available 4-channel electroencephalography (EEG) headset to access flow in game users and collect data from 22 participants while they played a video game. They manually computed 36 features and reported an accuracy of 50.1% with a subject-independent classification approach and 66.4% with a subject-dependent classification approach using a Support Vector Machine (SVM) classifier. However, they mentioned that new feature extraction and classification methods could improve their results.

Knierim et al. [35] analyzed electrocardiography (ECG) recordings to investigate the potential of predicting flow in the field through personalized models by collecting reports and ECG data from an office worker over two (2) weeks. They reported a mean absolute error of 1.18 for LASSO regression and an F1 score of 65% for binary Random Forest (RF) classification. However, for an adequate generalized model, the model should be tested with data from multiple subjects in a subject-independent manner, where data from test subjects should not be used to train the model.

Harmat et al. [36] investigated how psychological states change under different experimental conditions and what associations existed between self-reported psychological flow and physiological measures. For the experiments, they collected data from 77 participants who played Tetris under three experimental conditions (i.e., easy, optimal, and difficult). Physiological recordings were made continuously during all experimental conditions using ECG, respiration, and functional near-infrared spectroscopy (fNIRS). Statistica 12, an analytics software package originally developed by StatSoft [37] – was used to measure psychological state (flow, concentration, attentional performance, arousal, and valence) under all experimental conditions. Associations between self-reported psychological flow and physiological measures were examined using a series of repeated measures in linear mixed model analyses. The subjective flow was found to be positively related to the respiratory depth, and higher depth at high flow was indicative of a more relaxed state.

Rissler et al. [38] performed physiological flow classification using ML classifiers in some laboratory (lab) and field experiments. They recorded ECG chest band data for lab and field activities and computed features using the Python HRV package. The classification between low-level and high-level flow was performed using various state-of-the-art classifiers, for example, C4.5, RF, Adaptive Boosting (AdaBoost), SVM, Naïve Bayes (NB), and Decision Tree (DT), for comparison. They reported 70% accuracy in fieldwork and 68% in lab work for binary classification.

Di Lascio et al. [23] also performed a physiological classification of flow for work activities by collecting data from 13 participants using the Empatica E4 wristband [39]. In addition to manually crafted features, they investigated deep feature learning to classify the physiological features into binary classes (i.e., low-level flow vs. high-level flow). To solve the class imbalance problem, they used the *SMOTE* algorithm [40] and an accuracy of 70.93% with 5-fold cross-validation was achieved with a Convolutional Neural Network (CNN) based late fusion technique. They also analyzed the effects of contextual information related to the type of activity, time of day, and day of the week on perceived flow. They found that type of activity is relevant contextual information, and it should be considered when detecting flow during work activities.

Interestingly, Maier et al. [24] conducted experiments to determine the optimal user experience based on the physiological responses captured by the Empatica E4 wristband [39]. The features were computed manually and with deep feature learning from data collected from 72

participants playing the *Tetris* game with different difficulty levels. Experiments were conducted with 5-fold cross-validation for binary (low flow, high flow) and ternary (boredom, stress, and flow) classification. They reported an accuracy of 67.50% for binary classification using CNN-based feature learning and an accuracy of 49.23% for ternary classification using feature engineering with an RF classifier.

In general, there are few studies [11,23,24,35,36,38] on this topic that we investigate. Most studies examined flow in the context of video games, while others examined flow in the context of work-related activities. However, we are the first to investigate how emotional information can be transferred to flow recognition in the presence of subjective labels. In addition, we investigate the use of multimodal sensors in the context of human flow experience recognition and develop a state-of-the-art ML model that learns flow and non-flow patterns from physiological responses and classifies them into the flow and non-flow classes. We also analyze and compare wearable devices to select the most relevant physiological signals for accurate human flow state classification and provide a comparative analysis of feature extraction approaches and ML algorithms to achieve optimal classification results.

3. Materials and methods

In this section, we present the aspects of the sensor modalities used for data acquisition, the process of data acquisition, and preprocessing of the datasets. We also discuss the experimental settings and evaluation metrics used in this study.

3.1. Datasets

We used two datasets in this study, namely the *Physiological Sense Flow (PhySF)* and DEAP [31]. The PhySF dataset was collected at the *Assessment of Physical and Psychological Signals Laboratory* (APPS Lab) at the University of Lübeck [41] with the help of psychology partners from the *Department of Psychology at the University of Lübeck*. All preliminary experiments were conducted using the PhySF dataset. For the experiments based on the transfer learning approach, the DEAP dataset was used as the source dataset and PhySF as the target dataset. Sections 3.1.1 and 3.1.2 describe each dataset in detail, respectively.

3.1.1. Emotion recognition dataset: DEAP

DEAP is a dataset for emotion analysis in which 32 participants (16 males and 16 females aged between 19 and 37 years, with an average age of 26.9 years) participated in the data collection process. Each participant was asked to watch 40 video clips. Each video clip contained a one-minute music extract to elicit various emotional states. Physiological signals comprising EEG channels (n = 32) and peripheral channels (n = 15) were recorded while the participants watched the videos. In addition, before the actual recording, baseline signals were recorded for two minutes from each subject while they relaxed and looked at a fixation cross on a screen. After watching the videos, each participant had to rate each video regarding arousal, valence, dominance, and liking. The level of arousal, valence, dominance, and liking was assessed using the *Self-Assessment Manikin Scale* (SAM) [42], yielding numerical values between 1 (very low) and 9 (very high) for each emotional dimension.

In this study, we used the pre-processed version of the data sampled at 128 Hz made available online by the authors of the DEAP dataset.[1] The dataset is available upon request (signed EULA) for educational purposes. For more information about the dataset, the sensor devices, and the data collection process, see [31,43].

3.1.2. Flow recognition dataset: PhySF

Twenty-six (26) healthy individuals (8 males and 18 females aged between 18–40 years) participated in the PhySF dataset[2] collection, whose demographic information is presented in Appendix A. The sample consisted of students and employees of the University of Lübeck. Participants were required to be at least 18 years old, speak German fluently, and not have any cardiovascular disease, non-correctable vision, or movement impairment. The measurements were conducted under regional sanitary regulations during the Covid-19 pandemic between June and December 2022. The data of one volunteer was excluded from the dataset due to missing values. As well as psychology students could receive 1.5-course credits for their participation. The study was conducted following the *Declaration of Helsinki* and approved by the *Institutional Review Board of the University of Lübeck (April 14, 2022; No. 22-112)*.

Subjects were asked to read and sign the informed consent form before the experiment. The experiment began with a 5-minute baseline recording, during which participants were shown a fixation cross sign (they were advised to relax during this time). They were then instructed to perform mathematical and reading tasks using a web application [44]. The mathematical task consisted of 150 simple arithmetic questions, e.g., $7765 + 19 =$ __; $5583 - 7 =$ __, etc. The reading task consisted of a short story of about 3000 words ("Die verborgene Seite der Medaille" [The hidden side of the coin] by Scavezzon [45]). Participants kept wearing the sensor devices during the whole data collection. As well as, after each task, the participants indicated if and at what point of the task they experienced flow. The sensor devices used for data acquisition in this study can be seen in Fig. 1 and are described as follows:

1. The *Emotiv Epoc X* [46] is a 14-channel EEG headset for scalable and context-aware exploration of human brain activity. It is placed on the scalp, registering the bioelectrical activity of specific brain regions. The electrodes were hydrated using a saline solution. The EEG data were acquired during the experiments at a sampling rate of 128 Hz.
2. The *Empatica E4* [39] is a wearable wristband consisting of photoplethysmography (PPG), infrared thermopile, and EDA sensors that provide measurements of blood volume pulse (BVP), HR, interbeat interval (IBI), skin temperature (TMP), and galvanic skin response (GSR). During the data collection, subjects wore this band on the non-dominant hand. The description of each sensor contained in *Empatica E4* is as follows:
    - *PPG*: This sensor measures BVP. However, it is also useful in extracting other valuable information, such as HR and IBI. It is sampled at 1 Hz.
    - *Infrared Thermopile*: This sensor records TMP. It is sampled at 5 Hz.
    - *EDA*: This sensor measures GSR, which is the change in the skin's electrical conductivity in response to sweat secretion. It is also sampled at 5 Hz.
3. Biosignalplux RespiBAN [47] is a wearable device with a wearable, flexible respiratory belt and a hub. Subjects wore the respiratory belt around the chest, at the rib cage level, and the hub was attached to the belt with the electrode connections facing forward. The respiratory belt consists of a respiratory sensor (Resp), while the mounted hub provides the possibility to connect other sensors such as electrooculography (EOG), electromyography (EMG), and ECG. The data of each sensor channel was sampled at 475 Hz. The description of each sensor is as follows:

[1] https://www.eecs.qmul.ac.uk/mmv/datasets/deap/ (last accessed: 05.05.2023).

[2] https://osf.io/hgj6p/?view_only=1a23513a3ff24eae9afb47bcaba9f4f2 (last accessed: 03.08.2023).

Fig. 1. List of wearable devices utilized to acquire the physiological measurement of the participants: (a) - Headband style Emotiv Epoc X with 14 electrodes on flexible plastic armbands, (b) - Empatica E4 wristband, and (c) - Biosignalplux RespiBAN wearable device.

- *Resp*: This sensor detects the expansion and contraction of the chest or abdomen and outputs the respiratory rate. It is worn with a comfortable and adjustable belt.
- *EOG* [48]: The biosignalsplux EOG sensor is designed for seamless EOG data acquisition. It consists of two electrodes that record electrical potentials in a chosen specific facial region relative to a reference electrode. In this study, we placed two EOG electrodes on the right and left sides of the outer canthi. A reference electrode was placed on the back of the left ear. This sensor provides additional information about the subject's gaze patterns.
- *ECG* [49]: This sensor consists of three electrodes that allow unobtrusive ECG data acquisition. In this study, the electrodes were placed on the subject's right upper pectoral, left upper pectoral, and left bottom thoracic cage. This sensor records the electrical impulses through the heart muscle and can be used to extract HR data and other ECG features.
- *EMG* [50]: This sensor also consists of three electrodes that can measure the electrical activity associated with muscle contractions and the corresponding neurons that control them. In this study, two EMG electrodes were placed at the upper end of the trapezius muscle of the non-dominant side of the subject, and a reference electrode was at point C7.

After performing each task, subjects were asked to put themselves back to the task they had just completed and label their data by answering questions listed in Appendix B. The proportion of self-reported flow and non-flow data for each subject is shown in Fig. 2.

3.2. Pre-processing

In general, ML algorithms' performances depend on the data quality. In case of insufficient, unnecessary, and irrelevant data, they may provide inaccurate and less understandable results. Therefore, data pre-processing is an essential step in the ML pipeline. The steps performed to pre-process the DEAP and PhySF datasets are described in Sections 3.2.1 and 3.2.2.

3.2.1. DEAP

The version of the DEAP dataset [43] downsampled to 128 Hz was used for this study. As a first step, similar sensor channel data (also present in the PhySF dataset) were carefully separated for use in our transfer learning-based experiments. This is because the shape of the input data must be the same for both tasks (i.e., source and target) to successfully transfer the learned weights from the source domain to the target domain. The selected sensor channels of both datasets are shown in Table 1. It is worth noting that experiments were also performed with randomly selected sensor channels from both datasets, but the results were non-significant. Subsequently, the separated data were segmented using a *Sliding Window Segmentation* (SWS) technique. The current method for selecting the optimal window size is empirical [51]. People usually test different window sizes and choose the one that maximizes the recognition system's performance [52]. Therefore, window sizes of 10, 30, and 60 s were tested with a stride (i.e., step size) of 50% of the window sizes. The best settings were found with a window length T and a step size ΔS of 10 and 5 s, respectively.

Table 1
List of selected sensor channels from the DEAP and PhySF datasets used for the transfer learning approach. The GSR-1 and Temp sensor channels of the DEAP dataset are equivalent to the EDA and TMP sensor channels of the PhySF dataset and record similar physiological measurements.

DEAP dataset		PhySF dataset	
Channel No.	Channel name	Channel No.	Channel name
2	AF3	1	AF3
3	F7	2	F7
4	F3	3	F3
6	FC5	4	FC5
7	T7	5	T7
11	P7	6	P7
15	O1	7	O1
17	O2	8	O2
20	P8	9	P8
24	T8	10	T8
25	FC6	11	FC6
27	F4	12	F4
28	F8	13	F8
29	AF4	14	AF4
41	GSR-1	16	EDA
45	Resp	20	Resp
47	Temp	17	TMP

3.2.2. PhySF

The PhySF dataset comprises n = 23 sensor channels (i.e., 14 Emotiv Epoc, 5 Empatica E4, and 4 RespiBAN channels). The details of each wearable device and sensor channel are presented in Section 3.1.2. The data from each device had a different sampling rate. In the first step, data from all sensor channels were synchronized at a target frequency of 128 Hz using linear interpolation to infer the missing data values at the target timestamps. In the second step, the resampled data were segmented using the SWS technique. Values for the segmented window length T were tested for $T \in \{10, 30, 60\}$ (in seconds), with a segmentation step size ΔS always set to 50% of T. The best performances were

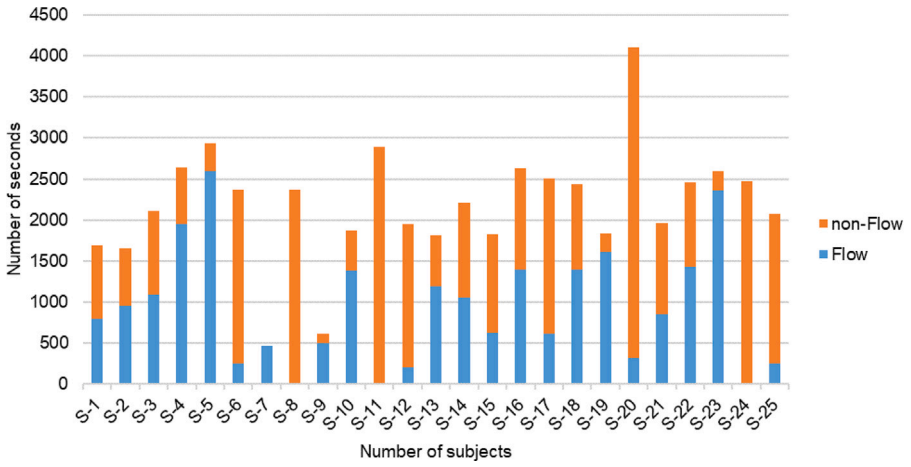

Fig. 2. The distribution of each subject's flow and non-flow class data in the PhySF dataset. *Y-axis:* represents the number of seconds of data for each class label. *X-axis:* represents the number of subjects in the PhySF dataset.

Table 2
List of hand-crafted features computed independently for each data segment of each sensor channel.

Hand-crafted features (HCF)		
Minimum	Mean	Maximum
Zero crossing	Standard deviation	20th percentile
50th percentile	80th percentile	Auto correlation
Skewness	Kurtosis	Interquartile
1st-order mean	2nd-order mean	Norm of 1st-order mean
Norm of 2nd-order mean	Spectral entropy	Spectral energy

Table 3
MLP architecture and hyperparameter values. The model was trained with the ADAM optimizer [55] using an initial learning rate of 5×10^{-4}.

Layer name	Neurons/Dropout rate	Activation
Batch Norm [56]	–	–
Dense	64	LeakyReLU (alpha = 0.2)
Dropout	0.50	–
Dense	32	LeakyReLU (alpha = 0.2)
Dropout	0.50	–
Dense	16	LeakyReLU (alpha = 0.2)
Dropout	0.50	–
Flatten	–	–
Dense	16	LeakyReLU (alpha = 0.2)
Dense	2	Softmax

obtained with $T = 10$ and $\Delta S = 5$ and are presented in Section 4. The results of the experiments with the other values of T are presented in Appendix C.

3.3. Feature extraction

In this study, we performed experiments using the following three feature extraction approaches:

- One feature engineering baseline that follows the traditional approach of manually computing simple statistical and frequency-related feature from the input time-series data.
- One feature learning baseline that follows training a DNN model end-to-end to learn features from raw input time-series data
- Our proposed approach follows transfer learning to enhance feature learning by transferring knowledge from emotion-related time-series data.

Each of the aforementioned approaches is described in more detail in the following subsections.

3.3.1. Feature engineering baseline

We calculated $F = 18$ manually crafted features on the PhySF dataset, also referred to as "hand-crafted features" (HCF). HCF characterize the data by computing simple statistical and frequency values on the physiological input signals or their power spectrum. All HCF used in this study are listed in Table 2. They were computed independently on each sensor channel for each data segment as suggested in [12,13], and then concatenated to form one $1D$ feature vector of size $F \times S = 18 \times 23 = 414$ representing a data segment.

3.3.2. Feature learning baselines

In this step, experiments were conducted using deep learning approaches. More specifically, a multilayer perceptron (MLP) and a CNN with a softmax classification layer were tested as feature learners. Since automatic optimization of DNN hyperparameters is still an obstacle so far [53,54], the trial-and-error method was used to find the best hyperparameters. The hyperparameters used in our experiments are listed in Tables 3 and 4, respectively.

3.3.3. Transfer learning-based approach

To enhance the performances obtained by feature learning approaches, we tested a method that consists of transferring DNN weights learned while training for a problem related to emotion recognition. We chose the DEAP dataset as the source domain because of its importance in the emotion recognition research field and since its data were acquired using similar sensor modalities as the PhySF dataset. More specifically, regarding the second point, the DEAP dataset consists of 48 channels (32 EEG and 16 others) that include 17 common channels with PhySF (i.e., AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4, GSR, TMP, and Resp). To maximize the chances of the transfer being successful, we used the subsets of the DEAP and PhySF datasets containing only the aforementioned sensor channels as source and target datasets, respectively. Fig. 3 illustrates the principle of our transfer learning approach.

The DEAP dataset contains emotion-related annotations, including arousal, valence, boredom, and liking ratings. However, following the standard practice of the literature [57–59], arousal and valence recognition are considered two separate binary classification problems between low and high arousal and valence for our source task. In the source task classification problem, we randomly split the data between training and testing, following a ratio of 70/30%. The DEAP dataset numerical ratings provided by the subjects were used to split the data into two classes using a value of 5 as the cutoff between the two classes (≤5 for low, >5 for high). Each classification problem, such as arousal (high vs. low) or valence (high vs. low), was used once separately

Table 4
CNN architecture and hyperparameter values. The model was trained with the ADAM optimizer [55] using an initial learning rate of 5×10^{-4} and a fixed dropout rate of 0.50.

Layer name	No. kernels (Units)	Kernel/Pool size	Stride	Activation
Batch Norm [56]	–	–	–	–
Convolutional	64	(2,1)	(1,1)	LeakyReLU (alpha = 0.2)
MaxPooling	–	(3,1)	–	–
Dropout	–	–	–	–
Convolutional	32	(2,1)	(1,1)	LeakyReLU (alpha = 0.2)
MaxPooling	–	(3,1)	–	–
Dropout	–	–	–	–
Convolutional	16	(2,1)	(1,1)	LeakyReLU (alpha = 0.2)
MaxPooling	–	(3,1)	–	–
Dropout	–	–	–	–
Flatten	–	–	–	–
Dense	16	–	–	LeakyReLU (alpha = 0.2)
Dense	2	–	–	Softmax

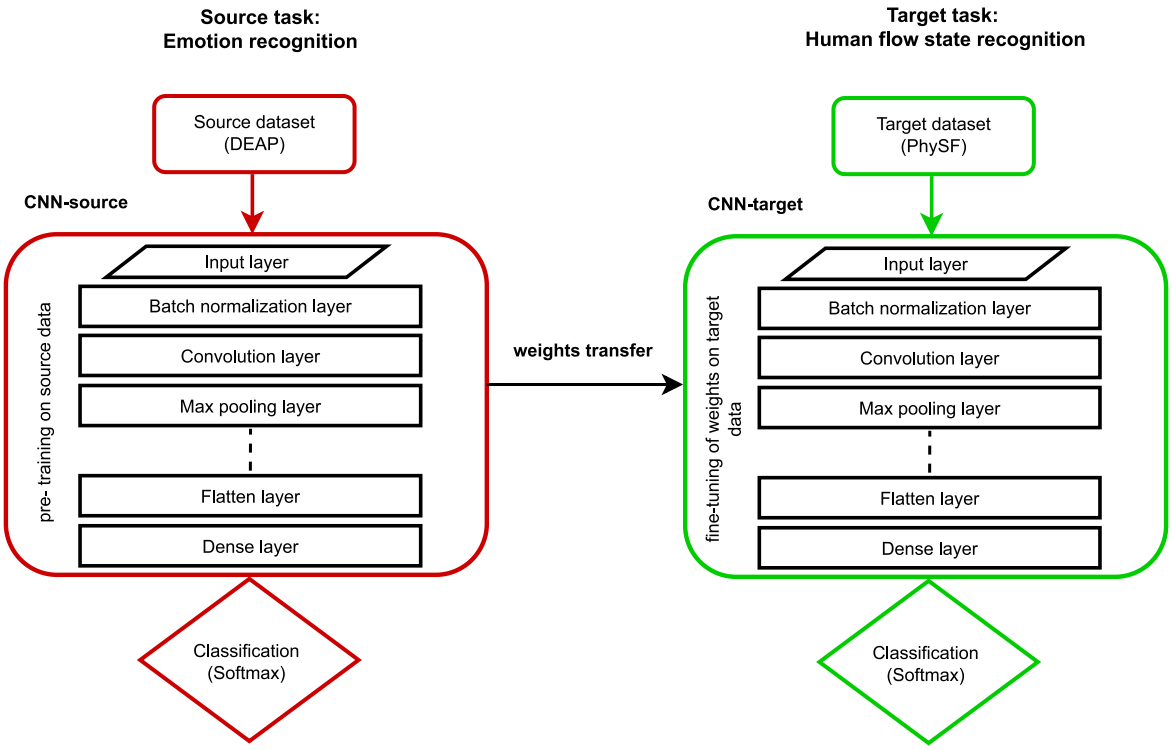

Fig. 3. Principle of our CNN-based transfer learning approach. In the first step, the CNN model is trained on the source dataset (i.e., DEAP) to solve the source task (emotion classification). In the second step, the weights learned at the first step are utilized to initialize the weights of the target model (CNN), which is then fine-tuned on the target dataset (i.e., PhySF) to recognize the human flow state.

as a source task to check which emotional dimension could lead to learning better flow detection features. A third classification problem with four classes (quadrants) involving both arousal and valence simultaneously (i.e., low arousal/low valence, low arousal/high valence, high arousal/high valence, and high arousal/low valence) was also tested as a source task.

To perform the knowledge transfer in order to enhance the feature learning, we chose the CNN model described in Section 3.3.2 as it showed higher performance in classifying flow and non-flow than the MLP baseline. The CNN model was appended to a softmax classification layer to obtain class estimations for the source task. We refer to this model as CNN-source. In the first step, the subset of the DEAP dataset that has 17 sensor channels in common with the PhySF dataset was used as the source dataset. CNN-source was trained end-to-end, and the weights and biases it learned were saved. Subsequently, the subset of the PhySF dataset that has 17 sensor channels in common with the DEAP dataset was used as the target dataset. The target task was to classify flow or non-flow into one of two categories. The weights and biases of the CNN-source were transferred to a CNN model trained to solve the target problem that we refer to as CNN-target. CNN-source and CNN-target share the same architecture, except for the softmax classification layer, which may vary in terms of the number of neurons which matches the number of classes of the source or target task considered. CNN-target was then fine-tuned in a supervised manner on the PhySF dataset. It is worth noting that in this experiment, no layer was frozen from the input layer to the flatten layer.

3.4. Classification

To obtain the best classification performances possible, we performed experiments with four different state-of-the-art classifiers. These include AdaBoost, RF, SVM, and Extreme Gradient Boosting (XGBoost) [12,60,61]. All classifiers were trained with the features extracted using the methods presented in Section 3.3.1. In the case of feature learning approaches (i.e., deep feature learning and transfer learning), the DNNs were trained following the standard procedure using a softmax classification layer.

As presented in Fig. 2, the PhySF dataset is highly imbalanced since some subjects did not report any occurrence of one of the two classes (e.g., the flow class is missing for subjects 8, 11, and 24, and the non-flow class is missing for subject 7). To consider this imbalance while obtaining features with a generalization capacity as large as possible, we trained all classifiers with a *Stratified-K-Fold - Cross Validation* (SKF-CV) for the target task. In our experiments, the subjects of the PhySF dataset were evenly split across $K = 5$ folds while ensuring that the number of subjects who predominantly reported being in flow and those not being in flow was mostly balanced within each fold. As a result, SKF-CV ensures that the following two conditions are verified during the training of the classifiers: both training and testing sets are balanced between the two classes, and the classifiers are trained in a subject-independent manner, i.e., evaluated on subjects who were not seen during the training process. For the PhySF dataset that contains $n = 25$ subjects, each classifier was trained on 20 subjects and evaluated on 5 subjects for each fold. The distribution of subjects in each fold and the flow and non-flow state data of each subject in seconds are presented in Table 11 of Appendix A.

3.5. Evaluation

To evaluate the classification performances of all models, the accuracy, sensitivity (also known as recall), and specificity were computed. Their formulas in terms of true positive (tp), true negative (tn), false positive (fp), and false negative (fn) are provided in Eqs. (1) to (3):

$$Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{1}$$

$$Sensitivity = Recall = \frac{t_p}{t_p + f_n} \tag{2}$$

$$Specificity = \frac{t_n}{t_n + f_p} \tag{3}$$

Because the accuracy may be biased by class imbalance, we also computed the average F1 score, also referred to as the macro Averaged F1 (AF1) score. The AF1 score is the average of all c classes F1 scores, where each class F1 score is the harmonic mean of the considered class precision and recall. Eqs. (4) to (6) provide the definitions of class precision, F1 score, and AF1 score. Recall (or sensitivity) is provided in Eq. (2).

$$Precision = \frac{t_p}{t_p + f_p} \tag{4}$$

$$F1score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

$$AF1score = \frac{1}{c} \sum_{i=1}^{c} \text{F1 score}_i \tag{6}$$

It should be noted that the training of an ML model is non-deterministic due to some randomness introduced during this process (e.g., random ordering of training examples, random weight initialization, etc.). This can therefore lead to some variance in the obtained results. To take this phenomenon into account, we reported the average results (values) of each metric obtained after training the model five times. All the results of our implemented approaches are presented in Section 4.

4. Experimental results

In this study, all models were implemented using Python 3.9.13. For the classifiers (i.e., AdaBoost, RF, SVM, and XGBoost) and DL models (i.e., MLP and CNN), the libraries scikit-learn, XGBoost, and Keras with a Tensorflow 2.10.0 backend were used. An *Adaptive Moment Estimation* (ADAM) with an initial learning rate of 5×10^{-4} was chosen as the optimizer for models based on DNN and trained with 50 epochs at a batch size of 16. The categorical cross entropy was used as the loss function for the DNN models. Since automatic optimization of hyperparameters is still an obstacle [53,54]. Therefore, all architectures and their associated parameters were found and optimized using the trial-and-error method.

Table 5
Flow and non-flow state recognition results with four different classifiers using feature engineering (i.e., HCF) for each wearable device data and the combination of all device's data with a window size (T) of 10, a step size (ΔS) of 5, and a sampling frequency of 128 Hz.

Classifier	Wearable device	Accuracy	AF1 score	Sensitivity	Specificity
AdaBoost	Emotiv	47.29	47.21	53.35	43.12
RF		45.07	40.68	21.92	60.99
SVM		42.66	38.45	20.27	58.05
XGBoost		49.74	47.53	35.84	59.30
AdaBoost	Empatica	56.59	55.28	48.51	62.14
RF		47.60	47.53	62.87	37.13
SVM		48.75	48.72	62.72	39.17
XGBoost		46.96	46.86	52.46	43.19
AdaBoost	RespiBAN	56.70	53.39	38.49	70.25
RF		55.76	45.81	15.86	83.13
SVM		55.92	38.86	03.80	91.67
XGBoost		51.20	45.89	24.45	69.54
AdaBoost	All	48.42	47.30	41.59	53.11
RF		48.78	46.29	33.51	59.25
SVM		49.58	47.27	35.20	59.44
XGBoost		42.54	42.20	42.88	42.31

AdaBoost: Adaptive Boosting; RF: Random Forest; SVM: Support Vector Machine; XGBoost: Extreme Gradient Boosting; AF1 Score: macro Averaged F1 score; Empatica: Empatica E4 wristband; RespiBAN: Biosignalsplux wearable device (Resp, ECG, EOG, EMG); Emotiv: Emotiv Epoc X – a 14 channel EEG headset.

Table 6
Flow and non-flow state recognition results using feature learning approach with MLP for each device data individually and for the combination of all device's data.

Wearable devices	Accuracy	AF1 score	Sensitivity	Specificity
Emotiv	64.39	63.83	60.87	67.01
Empatica	54.49	52.80	41.74	63.97
RespiBAN	57.09	56.50	53.35	59.86
All	**72.72**	**71.68**	**62.86**	**80.07**

4.1. Feature engineering baseline results

The 18 HCF listed in Table 2 were computed independently and used to train the four classifiers mentioned in Section 3.4, leading to a total of 414 features (i.e., Empatica: 18 × 5, RespiBAN: 18 × 4, and Emotiv: 18 × 14). To determine the best sensor modalities for flow classification, we also trained classifiers on features coming from each device separately. The evaluation metrics obtained for the best-performing segmentation parameters $T = 10$ and $\Delta S = 5$ s are provided in Table 5. The results obtained with the other values for the segmented parameters T and ΔS can be found in Appendix C.

As shown in Table 5, our feature engineering baseline results are mediocre for each device and for all device data that formed the basis for our feature learning approach. Furthermore, feature learning approaches have already performed better than feature engineering approaches in the past [13] in many application domains.

4.2. Feature learning baseline results

Similarly to the feature engineering baselines, we carried out experiments comparing the performances obtained with each device separately in addition to using all device data. The MLP and CNN baseline results are presented in Tables 6 and 7, respectively.

The results show that CNN-based feature learning is more efficient than MLP, with the latter achieving an accuracy of 73.63% when using data from multimodal sensing devices (Emotiv, Empatica, and RespiBAN) together. When used separately, the Emotiv device is more effective than Empatica and RespiBAN.

Table 7
Flow and non-flow state recognition results using feature learning approach with CNN for each device data individually and for the combination of all device's data.

Wearable devices	Accuracy	AF1 score	Sensitivity	Specificity
Emotiv	64.97	64.95	78.29	55.07
Empatica	60.08	57.51	41.82	73.67
RespiBAN	59.60	57.84	46.05	69.62
All	**73.63**	**72.70**	**64.02**	**80.78**

4.3. Transfer learning results

Since DNNs have performed well on large datasets, experiments were conducted with our implemented deep transfer learning approach to circumvent the data scarcity issue. We used the DEAP emotion recognition dataset as a source dataset and our PhySF as a target dataset, with 17 similar sensor channels from each. We hypothesize that transferring the learned weights of the emotion source dataset could improve flow recognition performance on PhySF. This is because, in the past, researchers have found a correlation between arousal and human flow experience [27,28]. The results of our transfer learning approach are presented in Table 8.

Improved results (i.e., the accuracy of 75.10% and AF1 score of 74.92%) were obtained with our transfer learning-based approach when the DEAP dataset was classified into high vs. low arousal. These results are about 5% higher than those without transfer learning using 17 sensor channels and also satisfy our hypothesis and confirm that better results can be obtained with the transfer learning approach when a related dataset is used as a source. Furthermore, the results show arousal is more critical than valence for flow detection.

4.4. Comparison to the literature

To the best of our knowledge, there is no publicly available dataset, and most of the literature work in this field is based on game user data or uses only the traditional feature engineering approaches. We searched the literature and could find only two studies that implemented the feature learning approach in this domain. The results of these studies are shown in Table 9. Furthermore, compared to the literature results reported in Table 9. This study achieved an improved accuracy of 73.63% and an AF1 score of 72.70% with the feature learning approach using CNN, an accuracy of 75.10%, and an AF1 score of 74.92% with the transfer learning approach using DEAP as a source dataset. The results of each of our implemented approaches (i.e., feature engineering, deep feature learning, and transfer learning) in the form of a confusion matrix are shown in Fig. 4(a–c). Moreover, our implemented models outperformed the results from the literature when trained in a subject-independent manner, while the work of Maier et al. [24] and Di Lascio et al. [23] were based on a subject-dependent strategy. It is also worth noting that we are the first to introduce transfer learning-based human flow recognition.

5. Discussion

The following points provide a detailed discussion of the aforementioned results:

- **Physiological signals for flow detection:** There are few studies on this topic in the literature, and most of them used a limited number of sensor channels compared to ours [11,23,24,62,63]. For example, Berta et al. [11] and Bartholomeyczik et al. [62] only analyzed EEG signals, Passalacqua et al. [63] performed experiments with only EDA data, Di Lascio et al. [23] and Maier et al. [24] considered only the investigation with Empatica E4 data for their experiments. Furthermore, they also reported mediocre performances of their models. Compared to the literature, this study used three sensor devices comprising physiological signals of 23 sensor channels (see Section 3.1.2 and Fig. 1) to detect human flow patterns. The effectiveness of each wearable device in terms of flow recognition with feature learning using CNN is illustrated in Fig. 5, which evidences correlations between flow and physiological variables contributed by each device (such as HR, EDA, and EEG [64–69]). Our multimodal approaches achieved notable performances (as shown in Tables 6, 7, and 8). They showed that it is feasible to measure the human flow state physiologically using ML with high accuracy using multimodal physiological sensors signals, and it is relevant because the flow has positive consequences [4–6].
- **Comparison to the state-of-the-art:** It is worth noting that most previous work in this domain is based on manual feature engineering approaches. To our best knowledge, only two past studies have used deep feature learning [23,24], but they also have some limitations. For example, Di Lascio et al. [23] mentioned that their work is based on something other than subject-independent cross-validation, which should be improved in future work to address the generalization capacity of the model. The work of Maier et al. [24] is also based on a subject-dependent cross-validation approach. Additionally, they also consider game users' data rather than work activities. We believe implementing a subject-independent cross-validation approach is mandatory to obtain a good generalized model. Thus, we performed experiments with subject-independent SKF-CV, and our models outperformed the previous results.
- **Feature engineering against feature learning:** We tested feature engineering and deep feature learning approaches to extract meaningful features from the raw physiological data. All our manually created features, or HCF, are listed in Table 2. However, HCF results were not convincing, as shown in Table 5. Thus, we used deep feature learning by implementing a fully connected neural network (i.e., MLP) and a CNN. It is worth noting that the deep feature learning approaches surpassed manual feature engineering in the literature [13]. This study also obtained notable results with CNN-based deep feature learning, as shown in Table 7. However, our transfer learning approach – which applies deep feature learning, outperformed our CNN-based feature learning when emotions-related information was included using weights transfer, as shown in Table 8.
- **Relation between flow and emotion:** Based on previous work that showed a relationship between flow and emotions [1,27], we investigated a method to transfer emotion-related information to flow recognition. More specifically, we used a CNN-based transfer learning approach with DEAP as the source dataset and PhySF as the target dataset. Our implemented approach outperformed the results of our CNN-based flow recognition and previous results in the literature when arousal classification was used as a source task, as shown in Table 8. Surprisingly, the transfer approach yielded slight improvement compared to the case without transfer whenever valence information was included in the target task. However, these results follow the literature findings as some researchers [2,28] suggested that flow is associated with affect and arousal. In a similar order of ideas, research by [27,36,70] suggested that moderate physiological arousal promotes flow, while boredom and stress (i.e., low and excessive physiological arousal) hinder it, as shown in Fig. 6. Regarding the relationship between flow and arousal, an experiment by Peifer et al. [27] provides evidence that even in the presence of a potentially negative stressor, flow can be experienced. The valence of emotional arousal, thus, can be subjectively reinterpreted when experiencing flow [71]. In line with this, the *Transactional Model of Stress and Flow* [8] proposes that a stressor that leads to a state of arousal can be interpreted as a challenge instead of a threat. A manageable challenge can result in the experience of flow [8]. This leads to the

Table 8
Flow and non-flow state recognition results using transfer learning approach with CNN. Two datasets were used, DEAP(17) and PhySF(17), which are the subsets of 17 common sensor channels, respectively taken from the DEAP and PhySF datasets. The source task consists of the classification of high arousal vs. low arousal, high valence vs. low valence, or quadrants (4 class-problem with low arousal/low valence, low arousal/high valence, high arousal/high valence, and high arousal/low valence) on the DEAP(17) dataset. The target task is the classification of flow vs. non-flow states on the PhySF(17) dataset.

Source task	Target task			
	Flow vs. non-flow classification on PhySF(17) dataset			
	Accuracy	AF1 score	Sensitivity	Specificity
None	70.09	70.01	89.36	56.13
High arousal vs. low arousal classification on DEAP(17)	**75.10**	**74.92**	**79.23**	**72.04**
High valence vs. low valence classification on DEAP(17)	70.30	70.24	88.80	56.88
Quadrants classification on DEAP(17)	71.53	71.47	90.34	57.89

PhySF(17): Physiological Sense Flow dataset of 17 common sensor channels (14-EEG + EDA + TMP + Resp); DEAP(17): DEAP dataset of 17 common sensor channels (14-EEG + GSR-1 + Temp + Resp); Temp and/or TMP: Skin temperature; EDA: Electrodermal activity; GSR: Galvanic skin response; Resp: Respiratory rate.

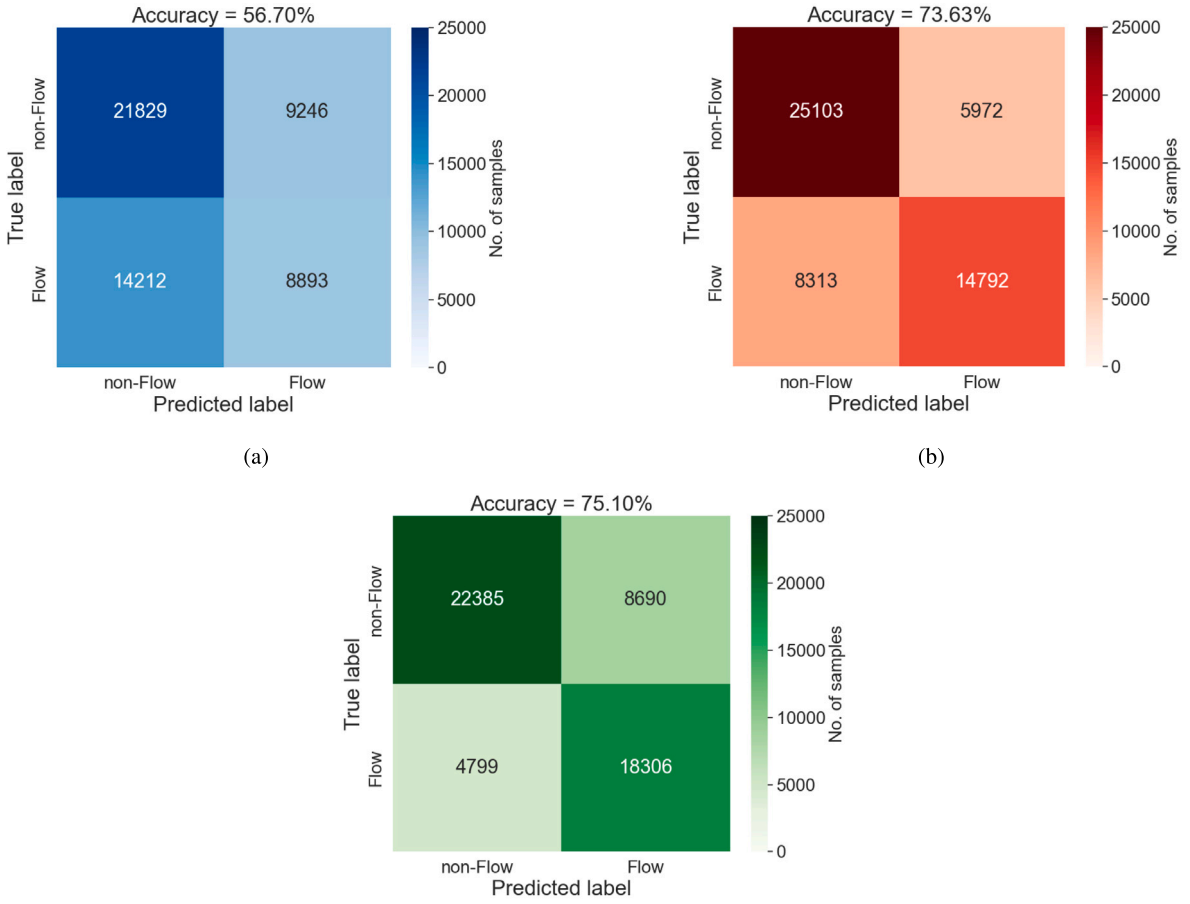

Fig. 4. Confusion matrices (a–c) for flow and non-flow recognition with three feature extraction approaches: (a) - feature engineering based on our hand-crafted features (HCF), (b) - feature learning using CNN, and (c) - transfer learning. In each figure (i.e., a–c), the top-left quadrant represents a true negative (tn) value, the top-right quadrant represents a false positive (fp) value, the bottom-left quadrant represents a false negative (fn) value, and the bottom-right quadrant represents a true positive (tp) value, respectively. All values (i.e., tp, tn, fn, fp) of the confusion matrices (a–c) are cumulative over 5 folds and the 5 runs of an experiment.

assumption that arousal is a reliable indicator for flow detection. In contrast, studies investigating emotional valence sensor data concerning flow showed inconsistent results, for example [8,72–74], and thus not yet be seen as reliable indicators for flow. Future research, however, should focus not only on arousal but also on valence.

6. Conclusions

In this work, we demonstrate the feasibility of using multimodal wearable devices for human flow experience recognition and the possibility of using emotion-related data to enhance human flow recognition. Our proposed multimodal system based on deep transfer learning

Table 9
Comparison of our implemented flow and non-flow recognition approaches with known results from the related literature.

Approach	Accuracy	AF1 score	Sensitivity	Specificity
Maier et al. [24]	67.50	–	–	–
Di Lascio et al. [23]	70.93	–	–	–
Our feature engineering	56.70	53.39	38.49	70.25
Our feature learning (CNN)	73.63	72.70	64.02	80.78
Our transfer learning	**75.10**	**74.92**	**79.23**	**72.04**

Our feature engineering: Best results using our hand-crafted features (HCF); Our feature learning: Best results of our CNN-based approach; Our transfer learning: Best results of our transfer learning approach.

Table 10
Demographic information of all the subjects (S1–S25) of the PhySF dataset.

Subject No.	Sex	Age (in years)	Subject No.	Sex	Age (in years)
S1	Male	35	S2	Female	25
S3	Female	23	S4	Male	37
S5	Female	27	S6	Female	24
S7	Male	25	S8	Male	27
S9	Male	28	S10	Male	24
S11	Female	27	S12	Female	20
S13	Female	19	S14	Female	18
S15	Male	28	S16	Female	24
S17	Female	40	S18	Female	20
S19	Female	21	S20	Female	22
S21	Female	24	S22	Female	23
S23	Male	19	S24	Female	19
S25	Female	19	–	–	–

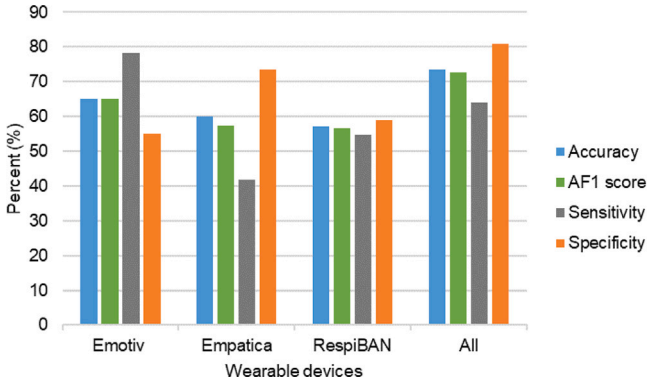

Fig. 5. Performance comparison of sensor devices based on accuracy, macro Averaged F1 (AF1) score, sensitivity, and specificity using feature learning when CNN model trained using the physiological data of Emotiv, Empatica, RespiBAN, and all devices together. Emotiv: Emotiv Epoc X 14 channel EEG headset; Empatica: Empatica E4 wristband; RespiBAN: RespiBAN professional device, including ECG, EMG, and EOG sensors.

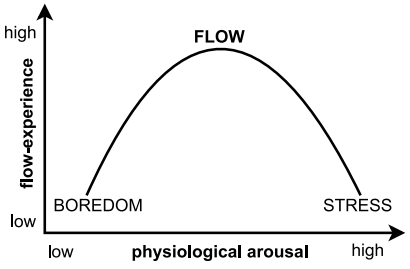

Fig. 6. Relationship between flow experience and physiological arousal. *Y-axis:* represents flow experience (low-high). *X-axis:* represents physiological arousal (low-high). Boredom and stress, which represent low and excessive physiological arousal, hinder the flow state, while moderate physiological arousal promotes it [27].

between the DEAP and our PhySF datasets discriminates between flow and non-flow states with an accuracy of 75.10% and an AF1 score of 74.92% in a subject-independent SKF-CV configuration. The results of this study lead to the following conclusions: Firstly, it is possible to achieve good flow recognition performance with multimodal physiological signals. Secondly, extracting emotion-based information related to flow and using it to enhance flow recognition performances is also feasible with transfer learning approaches which can also help to circumvent the data scarcity issue in this domain. Lastly, feature learning approaches could perform better in the context of flow recognition than feature engineering approaches. However, a limitation of this study is the limited dataset, consisting of only 25 participants. Future studies conducted with more subjects are necessary to validate our results further. In the future, we would like to collect more data related to the collaborative work environment [75,76] (i.e., team flow) and test the feasibility of our approach to identify team flow. We also wanted to investigate further research using only Empatica E4, as other devices, especially Emotiv Epoc X, are intrusive and not easy to use in practice.

CRediT authorship contribution statement

Muhammad Tausif Irshad: Conceptualization, Methodology, Investigation, Formal analysis, Software, Writing – original draft. **Frédéric Li:** Formal analysis, Validation, Writing – review & editing. **Muhammad Adeel Nisar:** Methodology, Formal analysis, Validation. **Xinyu Huang:** Investigation, Formal analysis, Validation. **Martje Buss:** Investigation, Validation, Writing – review & editing. **Leonie Kloep:** Validation, Writing – review & editing. **Corinna Peifer:** Funding acquisition, Investigation, Project administration. **Barbara Kozusznik:** Funding acquisition, Resources. **Anita Pollak:** Data curation, Resources. **Adrian Pyszka:** Validation, Data curation. **Olaf Flak:** Resources, Validation. **Marcin Grzegorzek:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition.

Declaration of competing interest

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

Funding

Research activities leading to this publication have been financially supported by the Narodowe Centrum Nauki (NCN), Poland, and the Deutsche Forschungsgemeinschaft (DFG), Germany, within the grant V-T-Flow "Team Flow and Team Effectiveness in Virtual Teams" (NCN: 2020/39/G/HS6/02124; DFG: 465142069).

Institutional review board statement

The study was conducted in accordance with the Declaration of Helsinki. The study was approved by the Institutional Review Board of the University of Lübeck, Germany (April 14, 2022; No. 22-112).

Informed consent statement

Informed consent was obtained from all subjects who participated in the study.

Appendix A. Demographic and cross-validation information

This appendix provides the demographical information (such as age and sex) of all (25) subjects of the PhySF dataset, as shown in Table 10. It also provides each subject's flow and non-flow state data (in seconds) and the distribution of the subjects in folds for SKF-CV, as shown in Table 11.

Table 11
Distribution of 25 subjects (S1–S25) to 5 different folds and the measurement of flow and non-flow data (in seconds) of each subject of the PhySF dataset.

Fold index	Subjects				
1	S1 (792 : 896)	S6 (254 : 2119)	S11 (0 : 2893)	S16 (1394 : 1235)	S21 (846 : 1119)
2	S2 (952 : 709)	S7 (467 : 0)	S12 (210 : 1745)	S17 (615 : 1889)	S22 (1429 : 1031)
3	S3 (1086 : 1027)	S8 (0 : 2374)	S13 (1189 : 627)	S18 (1397 : 1040)	S23 (2359 : 237)
4	S4 (1951 : 687)	S9 (502 : 106)	S14 (1058 : 1151)	S19 (1609 : 227)	S24 (0 : 2470)
5	S5 (2594 : 340)	S10 (1382 : 490)	S15 (619 : 1210)	S20 (316 : 3792)	S25 (246 : 1833)

Table 12
List of questions answered by participants at the end of each task to provide the labels of the flow and non-flow states. For question 1, only a single selection was possible. However, for question 2, multiple selections were possible. Furthermore, question 2 was only asked if the subjects answered 'yes' to question 1.

Question No.	Question	Options
1	Were you in the "Flow" during the task?	1. Yes 2. No
2	When were you in the "Flow" during the task?	1. First third of the task 2. The second third of the task 3. Last third of the task

Table 13
Flow and non-flow state recognition results with four different classifiers using feature engineering (i.e., HCF) for each wearable device data and the combination of all device's data with a window size $T \in \{30, 60\}$, a step size $\Delta S \in \{15, 30\}$, and a sampling frequency of 128 Hz.

Classifier	Wearable device	Window size (T)	Step size (ΔS)	Accuracy	AF1 score	Sensitivity	Specificity
AdaBoost	Emotiv	30	15	43.81	38.87	18.89	60.94
RF				43.40	41.36	30.39	52.34
SVM				40.04	37.44	24.14	50.97
XGBoost				41.59	36.49	16.26	59.00
AdaBoost	Empatica	30	15	50.10	49.99	55.83	46.18
RF				50.42	50.24	54.60	47.55
SVM				52.70	52.68	66.90	42.96
XGBoost				45.99	45.86	62.72	34.53
AdaBoost	RespiBAN	30	15	52.65	52.23	53.19	52.27
RF				55.50	49.13	24.73	76.60
SVM				54.58	41.19	08.44	86.23
XGBoost				55.63	39.49	04.89	90.42
AdaBoost	All	30	15	39.98	39.91	44.71	36.74
RF				45.97	42.45	26.11	59.60
SVM				47.48	46.13	38.87	53.39
XGBoost				45.55	41.94	25.34	59.43
AdaBoost	Emotiv	60	30	46.89	43.05	25.62	61.57
RF				44.67	38.02	14.60	65.41
SVM				44.26	39.41	19.55	61.32
XGBoost				46.94	43.58	27.60	60.29
AdaBoost	Empatica	60	30	46.63	45.45	75.32	26.92
RF				53.58	51.60	40.97	62.24
SVM				51.45	51.43	60.94	44.93
XGBoost				43.73	43.72	55.34	35.75
AdaBoost	RespiBAN	60	30	47.51	47.07	47.07	47.81
RF				56.58	48.51	20.87	81.12
SVM				54.35	44.09	14.12	81.99
XGBoost				54.30	44.11	14.10	81.39
AdaBoost	All	60	30	45.96	44.57	36.93	52.19
RF				53.57	49.33	30.20	69.67
SVM				48.24	47.24	42.26	52.36
XGBoost				51.97	46.61	24.87	70.63

Appendix B. Labeling of the recorded data

This appendix provides the questions asked to the study participants after each task in order to get the flow annotations of the PhySF dataset recordings. These questions are listed in Table 12.

Appendix C. Binary classification results using different window sizes and step sizes

This appendix provides the flow recognition results of each device data with feature engineering (i.e., HCF) using window sizes (T) $\in \{30, 60\}$ and step sizes (ΔS) $\in \{15, 30\}$ of four different ML classifiers, as shown in Table 13. The results with $T = 10$ and $\Delta S = 5$ are shown in Table 5.

References

[1] M. Csikszentmihalhi, Finding Flow: The Psychology of Engagement with Everyday Life, Hachette UK, 2020.
[2] M. Csikszentmihalyi, Beyond boredom and anxiety. san francisco: Josseybass, Well-being: Thefound. Hedonic Psychol. (1975) 134–154.
[3] S. Engeser, Theoretical integration and future lines of flow research, Adv. Flow Res. (2012) 187–199.
[4] M. Bassi, P. Steca, D. Monzani, A. Greco, A. Delle Fave, Personality and optimal experience in adolescence: Implications for well-being and development, J. Happiness Stud. 15 (2014) 829–843.

[5] C. Peifer, C. Syrek, V. Ostwald, E. Schuh, C.H. Antoni, Thieves of flow: how unfinished tasks at work are related to flow experience and wellbeing, J. Happiness Stud. 21 (2020) 1641–1660.

[6] R. Maeran, F. Cangiano, Flow experience and job characteristics: Analyzing the role of flow in job satisfaction, TPM-Test. Psychom. Methodol. Appl. Psychol. 20 (1) (2013) 13–26.

[7] C. Peifer, G. Wolters, Flow in the context of work, in: Advances in Flow Research, Springer, 2021, pp. 287–321.

[8] C. Peifer, J. Tan, The psychophysiology of flow experience, in: Advances in Flow Research, Springer, 2021, pp. 191–230.

[9] K. Nielsen, B. Cleal, Predicting flow at work: Investigating the activities and job characteristics that predict flow states at work, J. Occup. Health Psychol. 15 (2) (2010) 180.

[10] Y. Tezuka, N. Murayama, Y. Morioka, N. Suzuki, The influence of answer to the self-report scale on cardiovascular recovery, Int. J. Psychophysiol. 2 (94) (2014) 246.

[11] R. Berta, F. Bellotti, A. De Gloria, D. Pranantha, C. Schatten, Electroencephalogram and physiological signal analysis for assessing flow in games, IEEE Trans. Comput. Intell. AI Games 5 (2) (2013) 164–175.

[12] M.T. Irshad, M.A. Nisar, X. Huang, J. Hartz, O. Flak, F. Li, P. Gouverneur, A. Piet, K.M. Oltmanns, M. Grzegorzek, SenseHunger: Machine learning approach to hunger detection using wearable sensors, Sensors 22 (20) (2022) 7711.

[13] F. Li, K. Shirahama, M.A. Nisar, L. Köping, M. Grzegorzek, Comparison of feature learning methods for human activity recognition using wearable sensors, Sensors 18 (2) (2018) 679.

[14] X. Huang, K. Shirahama, F. Li, M. Grzegorzek, Sleep stage classification for child patients using DeConvolutional Neural Network, Artif. Intell. Med. 110 (2020) 101981.

[15] M.T. Irshad, M.A. Nisar, P. Gouverneur, M. Rapp, M. Grzegorzek, Ai approaches towards Prechtl's assessment of general movements: A systematic literature review, Sensors 20 (18) (2020) 5321.

[16] X. Huang, K. Shirahama, M.T. Irshad, M.A. Nisar, A. Piet, M. Grzegorzek, Sleep stage classification in children using self-attention and Gaussian noise data augmentation, Sensors 23 (7) (2023) 3446.

[17] N. Shahid, T. Rappon, W. Berta, Applications of artificial neural networks in health care organizational decision-making: A scoping review, PLoS One 14 (2) (2019) e0212356.

[18] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, Springer, 2014, pp. 818–833.

[19] C. Peifer, A. Kluge, N. Rummel, D. Kolossa, Fostering flow experience in HCI to enhance and allocate human energy, in: International Conference on Human-Computer Interaction, Springer, 2020, pp. 204–220.

[20] A.S. BaHammam, M.W. Chee, Publicly available health research datasets: opportunities and responsibilities, Nat. Sci. Sleep (2022) 1709–1712.

[21] D. Brickley, M. Burgess, N. Noy, Google Dataset Search: Building a search engine for datasets in an open web ecosystem, in: The World Wide Web Conference, 2019, pp. 1365–1375.

[22] Y. Fujiki, K. Kazakos, C. Puri, P. Buddharaju, I. Pavlidis, J. Levine, NEAT-o-Games: blending physical activity and fun in the daily routine, Comput. Entertain. (CIE) 6 (2) (2008) 1–22.

[23] E. Di Lascio, S. Gashi, M.E. Debus, S. Santini, Automatic recognition of flow during work activities using context and physiological signals, in: 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2021, pp. 1–8.

[24] M. Maier, D. Elsner, C. Marouane, M. Zehnle, C. Fuchs, DeepFlow: Detecting optimal user experience from physiological data using deep neural networks., in: AAMAS, 2019, pp. 2108–2110.

[25] L. Shao, Z. Cai, L. Liu, K. Lu, Performance evaluation of deep feature learning for RGB-D image/video classification, Inform. Sci. 385 (2017) 266–283.

[26] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.

[27] C. Peifer, A. Schulz, H. Schächinger, N. Baumann, C.H. Antoni, The relation of flow-experience and physiological arousal under stress—can u shape it? J. Exp. Soc. Psychol. 53 (2014) 62–69.

[28] Y. Tian, Y. Bian, P. Han, P. Wang, F. Gao, Y. Chen, Physiological signal analysis for evaluating flow during playing of computer games of varying difficulty, Front. Psychol. 8 (2017) 1121.

[29] M. Csíkszentmihályi, F. Massimini, M. Carli, The monitoring of optimal experience: A tool for psychiatric rehabilitation, J. Nervous Ment. Dis. 175 (1987) 545–549.

[30] C.E. Izard, C.E. Izard, Differential emotions theory, Hum. Emot. (1977) 43–66.

[31] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, Deap: A database for emotion analysis; using physiological signals, IEEE Trans. Affect. Comput. 3 (1) (2011) 18–31.

[32] P.A. Kragel, K.S. LaBar, Decoding the nature of emotion in the brain, Trends Cogn. Sci. 20 (6) (2016) 444–455.

[33] C. Li, Y. Hou, R. Song, J. Cheng, Y. Liu, X. Chen, Multi-channel EEG-based emotion recognition in the presence of noisy labels, Sci. China Inf. Sci. 65 (4) (2022) 140405.

[34] N. Takashima, F. Li, M. Grzegorzek, K. Shirahama, Embedding-based music emotion recognition using composite loss, IEEE Access (2023).

[35] M.T. Knierim, V. Pieper, M. Schemmer, N. Loewe, P. Reali, Predicting in-field flow experiences over two weeks from ECG data: A case study, in: NeuroIS Retreat, Springer, 2021, pp. 96–102.

[36] L. Harmat, Ö. de Manzano, T. Theorell, L. Högman, H. Fischer, F. Ullén, Physiological correlates of the flow experience during computer game playing, Int. J. Psychophysiol. 97 (1) (2015) 1–7.

[37] Statistica 12, 2023, https://www.statsoft.de/de/home [Online; accessed 31-Mar-2023].

[38] R. Rissler, M. Nadj, M.X. Li, N. Loewe, M.T. Knierim, A. Maedche, To be or not to be in flow at work: physiological classification of flow using machine learning, IEEE Trans. Affect. Comput. (2020).

[39] Empatica E4 wristband, 2023, https://www.empatica.com/research/e4/ [Online; accessed 25-Jan-2023].

[40] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.

[41] Apps lab, 2023, https://www.imi.uni-luebeck.de/forschung/p44-apps-lab.html [Online; accessed 13-Mar-2023].

[42] L. McEvoy, M. Smith, A. Gevins, Test–retest reliability of cognitive EEG, Clin. Neurophysiol. 111 (3) (2000) 457–463.

[43] DEAP: A dataset for emotion analysis using EEG, physiological, and video signals, 2023, https://www.eecs.qmul.ac.uk/mmv/datasets/deap/readme.html [Online; accessed 22-Feb-2023].

[44] Unipark, 2023, https://www.unipark.com/ [Online; accessed 31-Mar-2023].

[45] B. Scavezzon, Ein Album Voller Kurzgeschichten, Frankfurter Literaturverlag, 2010, pp. 49–58.

[46] Emotiv Epoc X: Scalable and contextual human brain wear — providing access to professional-grade brain data with an improved and easy-to-use design, 2023, https://www.emotiv.com/epoc-x/ [Online; accessed 25-Jan-2023].

[47] RespiBAN, 2023, https://plux.info/biosignalsplux-wearables/313-respiban-professional-820202407.html [Online; accessed 25-Jan-2023].

[48] Electrooculography (EOG), 2023, https://www.pluxbiosignals.com/products/electrooculography-eog-sensor-1 [Online; accessed 25-Jan-2023].

[49] Electrocardiogram (ECG), 2023, https://plux.info/sensors/277-electrocardiogram-ecg-820201203.html [Online; accessed 25-Jan-2023].

[50] Electromyography (EMG), 2023, https://plux.info/sensors/283-electromyography-emg-820201201.html [Online; accessed 25-Jan-2023].

[51] A. Dehghani, O. Sarbishei, T. Glatard, E. Shihab, A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors, Sensors 19 (22) (2019) 5026.

[52] G. Wang, Q. Li, L. Wang, W. Wang, M. Wu, T. Liu, Impact of sliding window length in indoor human motion modes and pose pattern recognition based on smartphone sensors, Sensors 18 (6) (2018) 1965.

[53] T. Yu, H. Zhu, Hyper-parameter optimization: A review of algorithms and applications, 2020, arXiv preprint arXiv:2003.05689.

[54] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13 (2) (2012).

[55] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[56] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, pmlr, 2015, pp. 448–456.

[57] F. Li, K. Shirahama, M.A. Nisar, X. Huang, M. Grzegorzek, Deep transfer learning for time series data based on sensor modality classification, Sensors 20 (15) (2020) 4271.

[58] J. Ma, H. Tang, W.-L. Zheng, B.-L. Lu, Emotion recognition using multimodal residual LSTM network, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 176–183.

[59] N. Ahmed, Z. Al Aghbari, S. Girija, A systematic survey on multimodal emotion recognition using learning algorithms, Intell. Syst. Appl. 17 (2023) 200171.

[60] T. Nguyen-Sy, J. Wakim, Q.-D. To, M.-N. Vu, T.-D. Nguyen, T.-T. Nguyen, Predicting the compressive strength of concrete from its compositions and age using the extreme gradient boosting method, Constr. Build. Mater. 260 (2020) 119757.

[61] Y. Zhao, X. Chen, J. Yin, Adaptive boosting-based computational model for predicting potential mirna-disease associations, Bioinformatics 35 (22) (2019) 4730–4738.

[62] K. Bartholomeyczik, M.T. Knierim, P. Nieken, J. Seitz, F. Stano, C. Weinhardt, Flow in knowledge work: An initial evaluation of flow psychophysiology across three cognitive tasks, in: Information Systems and Neuroscience: NeuroIS Retreat 2022, Springer, 2022, pp. 23–33.

[63] M. Passalacqua, R. Morin, S. Sénécal, L.E. Nacke, P.-M. Léger, Demystifying the first-time experience of mobile games: The presence of a tutorial has a positive impact on non-expert players' flow and continuous-use intentions, Multimodal Technol. Interact. 4 (3) (2020) 41.

[64] A.-M. Brouwer, M.A. Hogervorst, J.B. Van Erp, T. Heffelaar, P.H. Zimmerman, R. Oostenveld, Estimating workload using EEG spectral power and ERPs in the n-back task, J. Neural Eng. 9 (4) (2012) 045008.

[65] B. Mehler, B. Reimer, J.F. Coughlin, J.A. Dusek, Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers, Transp. Res. Rec. 2138 (1) (2009) 6–12.
[66] M.K. Karavidas, P.M. Lehrer, S.-E. Lu, E. Vaschillo, B. Vaschillo, A. Cheng, The effects of workload on respiratory variables in simulated flight: a preliminary study, Biol. Psychol. 84 (1) (2010) 157–160.
[67] B. Reimer, B. Mehler, The impact of cognitive workload on physiological arousal in young adult drivers: a field study and simulation validation, Ergonomics 54 (10) (2011) 932–942.
[68] O. Kohlisch, F. Schaefer, Physiological changes during computer tasks: responses to mental load or to motor demands? Ergonomics 39 (2) (1996) 213–224.
[69] J. Vogt, T. Hagemann, M. Kastner, The impact of workload on heart rate and blood pressure in en-route and tower air traffic control, J. Psychophysiol. 20 (4) (2006) 297–314.
[70] D.M. Bressler, A.M. Bodzin, A mixed methods assessment of students' flow experiences during a mobile augmented reality science game, J. Comput. Assist. Learn. 29 (6) (2013) 505–517.
[71] E.J. Donner, M. Csikszentmihalyi, Transforming stress to flow, Exec. Excell. 9 (1992) 16.
[72] Ö. De Manzano, T. Theorell, L. Harmat, F. Ullén, The psychophysiology of flow during piano playing, Emotion 10 (3) (2010) 301.
[73] J.M. Kivikangas, et al., Psychophysiology of flow experience: An explorative study, 2006.
[74] L.E. Nacke, C.A. Lindley, Affective ludology, flow and immersion in a first-person shooter: Measurement of player experience, 2010, arXiv preprint arXiv: 1004.0248.
[75] C. Peifer, A. Pollak, O. Flak, A. Pyszka, M.A. Nisar, M.T. Irshad, M. Grzegorzek, B. Kordyaka, B. Kożusznik, The symphony of team flow in virtual teams. Using artificial intelligence for its recognition and promotion, Front. Psychol. 12 (2021) 697093.
[76] F. Pels, J. Kleinert, Perspectives on group flow: Existing theoretical approaches and the development of the integrative group flow theory, Group Dyn.: Theory Res. Pract. (2022).