

Article

# Online Learning for Wearable EEG-Based Emotion Classification

Sidratul Moontaha <sup>\*,†</sup> , Franziska Elisabeth Friederike Schumann <sup>†</sup> and Bert Arnrich <sup>\*,†</sup> 

Digital Health—Connected Healthcare, Hasso Plattner Institute, University of Potsdam, 14482 Potsdam, Germany

\* Correspondence: sidratul.moontaha@hpi.de (S.M.); bert.arnrich@hpi.de (B.A.)

† These authors contributed equally to this work.

**Abstract:** Giving emotional intelligence to machines can facilitate the early detection and prediction of mental diseases and symptoms. Electroencephalography (EEG)-based emotion recognition is widely applied because it measures electrical correlates directly from the brain rather than indirect measurement of other physiological responses initiated by the brain. Therefore, we used non-invasive and portable EEG sensors to develop a real-time emotion classification pipeline. The pipeline trains different binary classifiers for Valence and Arousal dimensions from an incoming EEG data stream achieving a 23.9% (Arousal) and 25.8% (Valence) higher F1-Score on the state-of-art AMIGOS dataset than previous work. Afterward, the pipeline was applied to the curated dataset from 15 participants using two consumer-grade EEG devices while watching 16 short emotional videos in a controlled environment. Mean F1-Scores of 87% (Arousal) and 82% (Valence) were achieved for an immediate label setting. Additionally, the pipeline proved to be fast enough to achieve predictions in real-time in a live scenario with delayed labels while continuously being updated. The significant discrepancy from the readily available labels on the classification scores leads to future work to include more data. Thereafter, the pipeline is ready to be used for real-time applications of emotion classification.

**Keywords:** online learning; real-time; emotion classification; AMIGOS dataset; wearable EEG (muse and neurosky crown); psychopy experiments



**Citation:** Moontaha, S.; Schumann, F.E.F.; Arnrich, B. Online Learning for Wearable EEG-Based Emotion Classification. *Sensors* **2023**, *23*, 2387. <https://doi.org/10.3390/s23052387>

Academic Editors: Mariusz Szwoch and Agata Kolakowska

Received: 5 December 2022

Revised: 9 February 2023

Accepted: 13 February 2023

Published: 21 February 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Emotions play a crucial role in human communication and cognition, which makes comprehending them significant to understanding human behavior [1]. The field of affective computing strives to give emotional intelligence to machines that can recognize and interpret human affects [1,2], offering exciting possibilities for education, entertainment, and healthcare. Early detection and prediction of (mental) diseases or their symptoms can be facilitated, since specific emotional and affective states are often indicators thereof [3]. Moreover, long-term stress is one of today's significant factors causing health problems, including high blood pressure, cardiac diseases, and anxiety [4]. Notably, some patients with epilepsy (PWE) report premonitory symptoms or auras as specific affective states, stress, or mood changes, enabling them to predict an oncoming seizure [5]. This association of premonitory symptoms and seizure counts has been analyzed and validated from patient-reported diaries [6], and non-pharmacological interventions were proven to reduce the seizure rate [7]. However, many PWE cannot consistently identify their prodromal symptoms, and many do not perceive prodromes [8], emphasizing the necessity of objective prediction of epileptic seizures. In a previous work, the authors proposed developing a system to predict seizures by continuously monitoring their affective states [9]. Therefore, identifying pre-ictal states by measuring and predicting affective states in real-time through neurophysiological data could aid in finding pre-emptive therapies for PWE. That would be incredibly beneficial, especially to people with drug-resistant epilepsy, and would improve their quality of life [3,8]. Consequently, emotion detection in this paper is motivated by the idea that allowing computers to perceive and understand human emotions could

improve human–computer interactions (HCI) and enhance their ability to make decisions by adapting their reactions accordingly.

Since emotional reactions are seemingly subjective experiences, neurophysiological biomarkers, such as heart rate, respiration, or brain activity [10,11], are inevitable. Additionally, for continuous monitoring of affective states and thus detecting or predicting stress-related events reliably, low-cost, consumer-grade devices rather than expensive and immobile hospital equipment would be more meaningful [12]. This is an important area of interest in cognitive science and affective computing, with use cases varying from designing brain–computer interfaces [13,14] to improving healthcare for patients suffering from neurological disorders [15,16]. Among these, electroencephalography (EEG) has proven to be an accurate and reliable modality without needing external annotation [17,18]. Since clinical EEG is the gold standard for epileptic seizure detection [19], utilizing EEG-based emotion classification in detection systems could potentially predict seizures by knowing the affective states. Moreover, with recent advancements in wearable technology, consumer-grade EEG devices have become more accessible and reliable, opening possibilities for countless real-life applications. Wearable EEG devices such as the *Emotiv EPOC* headset or the *Muse S* headband have become quite popular tools in emotion recognition [20–22]. The *Muse S* headband has also been used for event-related potential (ERP) research [12] and for the challenge of affect recognition in particular. More specifically, *Muse S* has already been used in experimental setups to obtain EEG data from which the mental state (relaxed/concentrated/neutral) [13] and the emotional state (using the valence-arousal space) [23] could be reliably inferred through the use of a properly trained classifier.

### 1.1. Problem Statement

A challenging but essential step to identifying stress-related events or improving HCI in real-life settings is to recognize changes in peoples' affect by leveraging live data. The EEG-based emotion classifications mentioned in the literature have nearly exclusively been employed in traditional machine learning strategies, i.e., offline classifiers, and are often combined with complex data pre-processing techniques on static datasets [24,25]. Such cases expect the whole dataset, including labels, to be present for model training, unlike real application scenarios, where the data source is primarily a live data stream and classification results are required in real-time. Moreover, a live data stream from a non-stationary EEG data source presents challenges associated with, for example, the online arrival of data, the velocity of data, the data volume over time, and the dynamic nature of data [26]. The training and classification time for the offline models and the static behavior dealing with upcoming data streams makes real-time classification of affective states infeasible. Therefore, for live emotion classification, employing online learning by updating a pre-trained model continuously on new data is inevitable. Additionally, to ensure the usability of such a system in daily life, it is necessary to classify the data from a portable EEG device. Furthermore, the existing methodologies in the literature have primarily been developed and evaluated on curated EEG datasets, which need more reproducibility to apply to live data from real applications.

Regarding the mentioned problems, the primary goal of this paper is to answer the research question of how reliably an online classifier classifies emotion from state-of-the-art AMIGOS data [22] and how accurately the classifier can perform affective state prediction while curating EEG data from wearable EEG devices in live settings.

### 1.2. Key Contributions

Therefore, *firstly*, the key contribution of this paper is the establishment of a lightweight emotion classification pipeline that can classify a person's affective state based on an incoming EEG data stream in real-time, efficiently enough to be used in real applications, e.g., for seizure prediction. The developed pipeline leverages online learning to train subject-specific models on data streams by implementing binary classifiers for the affect dimensions: *Valence* and *Arousal*. The pipeline is validated by streaming an existing dataset

of established quality, AMIGOS, with better classification performance than state-of-the-art contributions.

*Secondly*, an experimental framework is developed, similar to the AMIGOS dataset, which can collect neurophysiological data from a wide range of commercially available EEG devices and show live prediction results of the subjects' affective states even when labels arrive with a delay. Data from 15 participants were captured by using two consumer-grade EEG devices.

*Thirdly*, the most novel contribution of this paper is to validate the pipeline on the curated dataset by wearable EEG devices in the first experiment with consistent classification performance on the AMIGOS dataset. Following this, live emotion prediction was performed successfully on an incoming data stream in the second experiment with delayed incoming labels.

The curated data from the experiments and metadata are accessible to the designated researchers as per the participants' consent. Therefore, the dataset is available upon request for scientific use via a contact form on Zenodo: <https://doi.org/10.5281/zenodo.7398263> (accessed on 18 February 2023). The Python code for loading the dataset and implementations of the developed pipeline have been made available on GitHub: <https://github.com/HPI-CH/EEGEMO> (accessed on 18 February 2023). The next section will explain the related works, material, and methods utilized within this paper following the results and discussion sections.

### 1.3. Related Work

Recent review papers from Dadebayev et al. [24] and Suhaimi et al. [25] mention several articles on EEG-based emotion recognition using various affect dimension scales, EEG devices, machine learning algorithms, and performance matrices. However, only a few research works have mentioned *real-time* emotion classification. Müller et al. [27] proposed an online linear discriminate analysis (LDA) classifier by utilizing spectral features in the EEG data obtained from a brain-computer interface (BCI) device. However, while demonstrating the application of real-time arousal monitoring, the researchers used data from one participant to train an offline LDA classifier. Moreover, the data acquisition took place utilizing a 128-channel BCI, and applying the methodology to the data from wearable EEG devices was mentioned in their future work, which already falls into the scope of this paper.

Liu et al. [28,29] showed promising results on emotion recognition using the fractal dimension (FD) model on a 14-channel EEG headset. The applied methodology in real-time application was reflected as computer avatars demonstrating a person's emotion based on live data. The fractal algorithm required 1024 samples at a time, obtained by the device's buffer function. The buffer function may not be useful for general applications and EEG devices. A follow-up work was proposed by Lan et al. [30,31], analyzing the stable features of the previously mentioned application. Hou et al. [32] further developed a meter to visualize the intensity of the felt emotion. However, the mentioned live emotion classification was based on a static model and was not updated during prediction. We propose a general emotion classification pipeline that deals with a selection of consumer-grade EEG devices and general applications, which can be validated on the published dataset from the researchers mentioned above [33].

Additionally, Javaid et al. [34] reported a 25.44% higher accuracy while switching from eight to three wet electrodes of an openBCI Kit while quantifying four basic emotions using an FD threshold-based classifier similar to [28]. However, the authors also mentioned incorporating an Emotiv EPOC device in a second session of their proposed method, but the follow-up research still needs to be reported.

Sarno et al. [35] used power features in the K-nearest neighbor (KNN) classifier to train 14-channel EEG data offline and predict binary and tertiary emotion in an online stage. An evaluation only of accuracy was reported, which is compatible with our results, given that the F1-Score reported in our paper is relatively high. Moreover, our

paper uses fewer electrodes and multiple algorithms and develops and publishes the data collection framework.

Very recently, Bajada et al. [36] built an emotion classification pipeline incorporating discrete wavelet transforms (DWT) features into a 3D convolutional neural network (3D CNN) and support vector machine (SVM) classifiers from pre-recorded and pre-trained data from the state-of-the-art DEAP [37] dataset. They proposed using the proposed algorithm for real-time monitoring because it maintains a high accuracy of up to 92%, reducing from 32 channels to 5 channels. Li et al. [38] addressed the challenge of when a model can see the data only once by leveraging cross-subject and cross-session data by implementing the *Fast Online Instance Transfer (FOIT)* algorithm. They validated their methodology on two state-of-the-art datasets: SEED [39] and SEED-IV [40]. However, these mentioned studies need to apply live incoming data streams similar to our approach, which pose challenges such as velocity, veracity, and concept drift. To the best of our knowledge, only Nandi et al. [41] have employed online learning to classify emotions from an EEG data stream from the DEAP dataset and proposed an application scenario in e-learning, but have yet to report undertaking any such live experiments. They compared the performance of different state-of-the-art online classifiers, such as adaptive random forest (ARF) [26] and Hoeffding adaptive tree (HAT) [42], against their own real-time emotion classification system (RECS) on the DEAP dataset.

Indeed, more research is needed on using online machine learning for emotion recognition.

Moreover, multi-modal labeled data for the classification of affective states have been made freely available through annotated affective databases, such as DEAP [37], DREAMER [43], ASCERTAIN [44], SAFE [33], and AMIGOS [22], which play a significant role in further enhancing the research of this field. They include diverse data from experimental setups using differing emotional stimuli such as music, videos, pictures, or cognitive load tasks in an isolated or social setting. Such databases enable the development and improvement of frameworks and model architectures with existing data of ensured quality. However, none of these published datasets include the data collection framework to be reused in curating the data from wearable EEG devices in live settings.

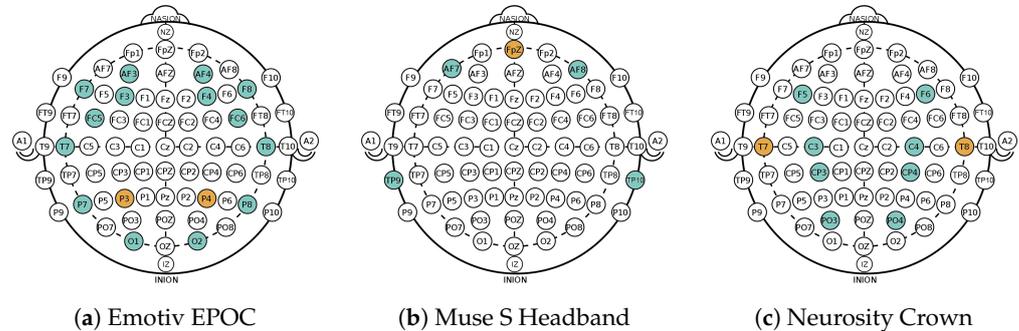
## 2. Materials and Methods

### 2.1. Dataset I: AMIGOS Dataset

The developed emotion classification pipeline in this paper was evaluated on the state-of-art dataset for affect, personality, and mood research on individuals and groups (AMIGOS) published by Miranda-Correa et al. [22], which is further referred to as *Dataset I*. Upon following the data receiving protocol, all data from the AMIGOS dataset that are used in this work stem from short video individual experiments where 40 healthy participants (13 female) aged between 21 and 40 (mean age 28.3) were asked to watch 16 videos from defined movie clips. The EEG data were recorded using the Emotiv EPOC headset (<https://www.emotiv.com/epoc-x/> (accessed on 20 February 2023)) with a sampling frequency of 128 Hz and a 14 bit resolution. This device records EEG data from 14 channels (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4) of the brain according to the 10–20 system as depicted in Figure 1a.

Additionally, Dataset I reports the Self-Assessment Manikin [45] with a scale from 1 to 9 as recording participants' affect ratings of the dimensions valence, arousal, and dominance. The participants were also asked to rate their familiarity with the videos and whether they liked them, and had to select at least one option from a list of basic emotions felt after watching each video. However, only the obtained valence and arousal ratings were considered the ground truth for this paper. Furthermore, the participants answered the Positive and Negative Affect Schedules (PANAS) [46] questionnaire at the beginning and end of the experiment; only one overall calculated PANAS score is reported. Pre-processed data files were used for classification, where the EEG data were down-sampled to 128 Hz, averaged to a common reference, and a band pass filter was applied from 4.0–45.0 Hz as described in the description on the website (<http://www.eecs.qmul.ac.uk/>

[mmv/datasets/amigos/readme.html](https://github.com/mmv/datasets/amigos/readme.html) (accessed on 20 February 2023)). The files containing electrocardiogram (ECG) and galvanic skin response (GSR) data were removed for the analysis of this paper.



**Figure 1.** Different electrode positions according to the international 10–20 system of the EEG devices used in Dataset I (a) and in Dataset II and III (b,c). Sensor locations are marked in blue, references are in orange.

To validate the established pipeline in daily life and live setup, two different datasets with individual experimental protocols, named *Dataset II* and *Dataset III*, were curated with the description of participants, data acquisition, and experimental protocols mentioned in the following sections.

### 2.2. Participants

For Dataset II, eleven participants were recruited (six females and five males) between the ages of 25 and 42 ( $\mu = 29.27$ ,  $\sigma = 5.41$  years). Data from two participants had to be discarded for further analysis. Dataset III was made with data from four participants (one female and three males) between the ages of 25 and 34 ( $\mu = 28.5$ ,  $\sigma = 3.5$  years). People who were pregnant, older than 65 years, and had taken part in one of the experiments were excluded from participation. All participants had normal or corrected vision and reported no history of neurological or mental illnesses or head injuries.

### 2.3. Data Acquisition

**Hardware:** During the experiments, two consumer-grade devices, *Muse S Headband* Gen 1 (<https://choosemuse.com/compare/> (accessed on 20 February 2023)) and *Neurosity Crown* (<https://neurosity.co/crown> (accessed on 20 February 2023)), were used to collect the EEG data from the participants, as depicted in Figure 2. Both devices operated with a sampling rate of 256 Hz and the EEG data were collected with four and eight channels, respectively.



(a) Muse S Headband

(b) Neurosity Crown

**Figure 2.** Two consumer-grade EEG devices with integrated electrodes used in the experiments.

According to the international 10–20 system [47], the channels on the Muse S Headband correspond to AF7, AF8, TP9, and TP10 (see Figure 1b), with a reference electrode at Fpz [12]. The channel locations of Neurosity Crown are C3, C4, CP3, CP4, F5, F6, PO3, and PO4, with reference sensors located at T7 and T8, as shown in Figure 1c. Using the Mind Monitor App (<https://mind-monitor.com/> (accessed on 20 February 2023)), the raw EEG data were streamed from Muse to a phone via Bluetooth. The app sends the

data to a laptop via the open sound control (OSC) protocol and the python-osc library (<https://pypi.org/project/python-osc/> (accessed on 20 February 2023)) on the receiving end. As the incoming data tuples from the Muse Monitor App did not include timestamps, they were added by the pipeline upon arrival of each sample. Similarly, the Crown uses the python-osc library to stream the raw EEG data to a laptop without enabling any pre-processing settings. In contrast to the Muse Headband, the Crown includes a timestamp when sending data.

**Software:** In this paper, the experiment was implemented using the software PsychoPy (v 2021.2.3) [48] in a way that guided the participants through instructions, questionnaires, and stimuli. The participants were allowed to go at their own pace by clicking on the “Next” (“Weiter” in German) button, as shown in the screenshots of PsychoPy in Figure 3.

#### 2.4. Stimuli Selection

Inducing specific emotional reactions is a challenge, even in a fully controlled experimental setting. Several datasets have tried to solve this issue with different modalities such as pictures [49–51], music [52,53], music videos [37,54,55], or combinations of them [56]. In this work, videos depicting short movie scenes were used as stimuli, based on the experimental setup of Dataset I [22]. Therefore, 16 short clips (51–150 s long,  $\mu = 86.7$  s,  $\sigma = 27.8$  s) depicting scenes from 15 different movies were used for emotion elicitation. Twelve of these videos stem from the DECAF dataset [54], and four movie scenes were taken from the MAHNOB-HCI [55] dataset. According to Miranda-Correa et al., these specific clips were chosen because they “lay further to the origin of the scale” than all other tested videos. This means they represent the most extreme ratings in their respective category according to the labels provided by 72 volunteers. The labels were provided in the two-dimensional plane spanned by the two dimensions *Valence* and *Arousal* according to Russell’s circumplex model of affect [57]. Valence, the dimension describing one’s level of pleasure, ranges from sad (unpleasant and stressed) to happy (pleasant and content), and Arousal ranges from sleepy (bored and inactive) to excited (alert and active). Therefore, the four category labels are High Arousal, Low Valence (HALV); High Arousal, High Valence (HAHV); Low Arousal, High Valence (LAHV); and Low Arousal, Low Valence (LALV). The selected movie scenes described in Table 1 are balanced between each of the valence–arousal space quadrants (HVHA, HVLA, LVHA, and LVLA). The video ID 19 corresponded to a scene from the movie *Gandhi*, which differs from the AMIGOS dataset but falls into the same LALV quadrant.

**Table 1.** The source movies of the videos used are listed per quadrant in the valence–arousal space. Video IDs are stated in parentheses, sources marked with a † were taken from the MAHNOB-HCI dataset [55]; all the others stem from DECAF [54]. In the category column, H, L, A, and V stand for high, low, arousal, and valence, respectively. This table has been adapted from Miranda-Correa et al. [22].

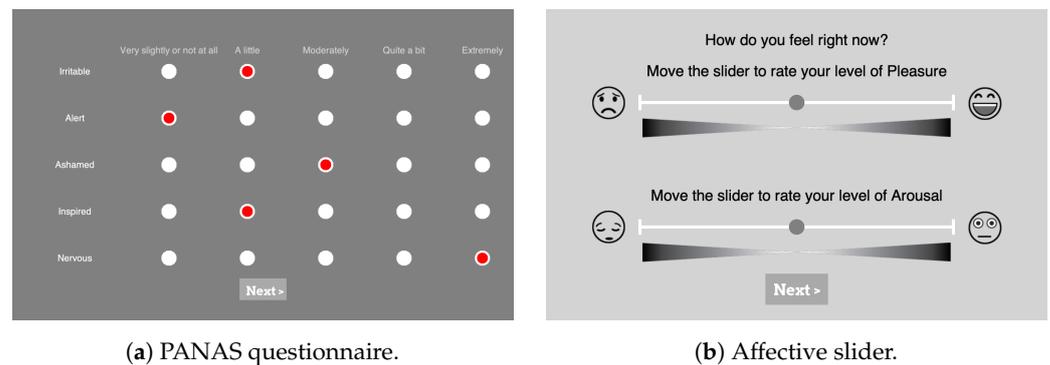
Category	Source Movie
HAHV	<i>Airplane</i> (4), <i>When Harry Met Sally</i> (5), <i>Hot Shots</i> (9), <i>Love Actually</i> (80) <sup>†</sup>
LAHV	<i>August Rush</i> (10), <i>Love Actually</i> (13), <i>House of Flying Daggers</i> (18), <i>Mr Beans’ Holiday</i> (58) <sup>†</sup>
LALV	<i>Gandhi</i> (19), <i>My Girl</i> (20), <i>My Bodyguard</i> (23), <i>The Thin Red Line</i> (138) <sup>†</sup>
HALV	<i>Silent Hill</i> (30) <sup>†</sup> , <i>Prestige</i> (31), <i>Pink Flamingos</i> (34), <i>Black Swan</i> (36)

#### 2.5. Behavioral Data

**PANAS:** During the experiments, participants were asked to assess their baseline levels of affect in the PANAS scale. As depicted in Figure 3a, in total 20 questions (10 questions from each of the Positive Affect (PA) and Negative Affect (NA) dimensions) were answered using a 5-point Likert scale with the options ranging from “very slightly or not at all” (1) to “extremely” (5). To see if the participants’ moods generally changed over the course of the

experiments, they were asked to answer the PANAS once at the beginning and once again at the end. For the German version of the PANAS questionnaire, the translation of Breyer and Bluemke [58] was used.

**Affect Self-Assessment:** The Affective Slider (AS) [59] was used in the experiment to capture participants' emotional self-assessment after presenting each stimulus, as depicted in the screenshot in Figure 3b (<https://github.com/albertobeta/AffectiveSlider> (accessed on 20 February 2023)). The AS is a digital self-reporting tool composed of two slider controls for the quick assessment of pleasure and arousal. The two sliders show emoticons at their ends to represent the extreme points of their respective scales, i.e., unhappy/happy for pleasure (valence) and sleepy/wide awake for arousal. For the experiments, AS was designed in a continuous normalized scale with a step size of 0.01 (i.e., a resolution of 100), and the order of the two sliders was randomized each time.



(a) PANAS questionnaire.

(b) Affective slider.

**Figure 3.** Screenshots from the PsychoPy [48] setup of self-assessment questions. (a) Partial PANAS questionnaire with five different levels represented by clickable radio buttons (in red) with the levels' explanation on top, (b) AS for valence displayed on top and the slider for arousal on the bottom.

**Familiarity:** The participants were asked to indicate their familiarity with each video on a discrete 5-point-scale ranging from “Have never seen this video before” (1) to “Know the video very well” (5). The PsychoPy slide with this question was always shown after the AS.

## 2.6. Dataset II

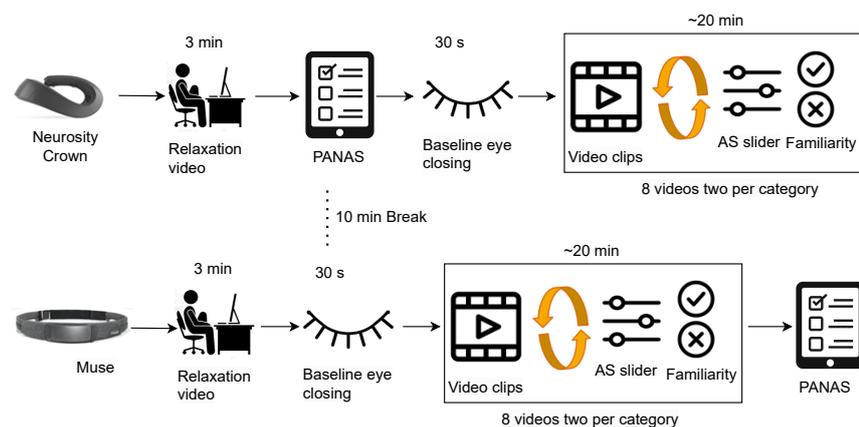
**Briefing Session:** In the beginning, each participant went through a pre-experimental briefing where the experimenter explained the study procedure and informed the participant of the experiment's duration, i.e., two parts of approximately 20 min each with a small intermediate break. The participant then received and read the data information sheet, filled out the personal information sheet, and signed the consent to participate. Personal information included age, nationality, biological sex, handedness (left- or right-handed), education level, and neurological- or mental health-related problems. The documents and the study platform (i.e., PsychoPy) were provided according to the participant's choice of study language between English and German. Afterward, the experimenter explained the three scales mentioned in and allowed the participant to accustom to the PsychoPy platform. This ensured the understanding of the different terms and scales used for the experiment without having to interrupt the experiment afterwards. The participant could refrain from participating at any moment during the experiment.

**Data Collection:** After the briefing, the experimenter put either the Muse headband or the Crown on the participant by a random choice. Putting headphones over the device, the participant was asked to refrain from strong movements, especially of the head. The experimenter then checked the incoming EEG data and let the participant begin the experiment. After greeting the participant with a welcome screen, a relaxation video was shown to the participant (<https://www.youtube.com/watch?v=S6jCd2hSVKA> (accessed on 20 February 2023)) for 3 min. They answered the PANAS questionnaire to rate their current mood and closed eyes for half a minute to get a baseline measure of EEG data. Afterwards, they were asked to initially rate the valence and arousal state with the AS.

Following this, an instruction about watching eight short videos was provided. Each of those was preceded by a video counter and followed by two questionnaires: the AS and the familiarity. The order of the videos and the order of the two sliders of AS were randomized over both parts of the experiments, fulfilling the condition that the labels of the videos are balanced. The first part of the experiment ended after watching eight videos and answering the corresponding questionnaire. The participant was allowed a short break after taking the EEG device and the headphones off.

In the second part of the experiment, the experimenter put the device that had not been used in the first part (Muse or Crown, respectively) and the headphones on the participant. Subsequently, the experimenter started the second part of the experiment, again after ensuring that the data collection was running smoothly. The participant followed the exact same protocol: watching the relaxation video, closing their eyes, and watching eight more movie scenes with the AS and familiarity questions in between. Lastly, they were asked for a final mood self-assessment via a second PANAS questionnaire to capture differences before and after the experiment. The experimental setup for curating Dataset II is depicted in Figure 4.

In this experiment, one PC (with a 2.4 to 3.0 GHz Dual Intel Core i5-6300U and 12 GB RAM) was used to present the stimuli and store EEG data to be used only after the experiment session.

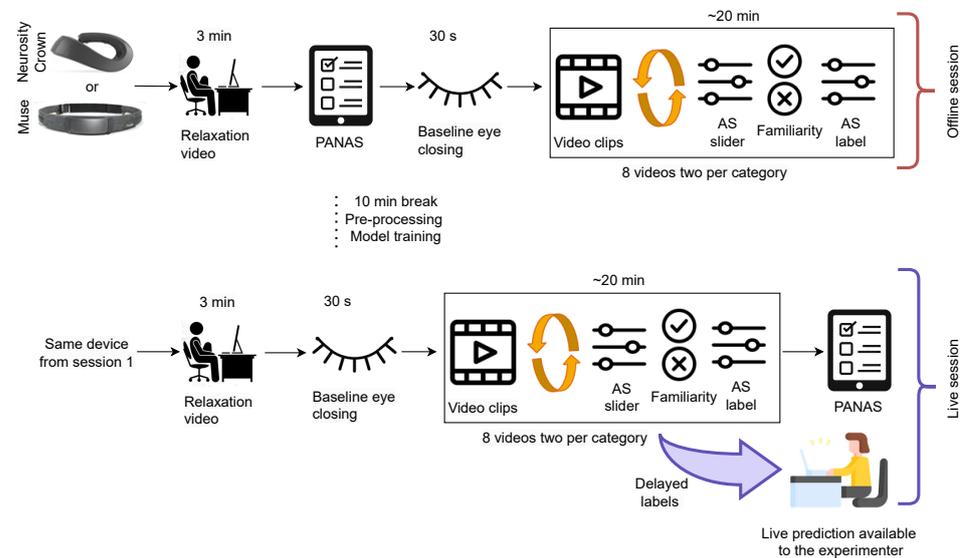


**Figure 4.** Experimental setup for curating Dataset II. The participants watched a relaxation video at the beginning and eight videos, two of each dimension category wearing one of the two devices. Between the eight videos, they answered AS slider and familiarity with the video.

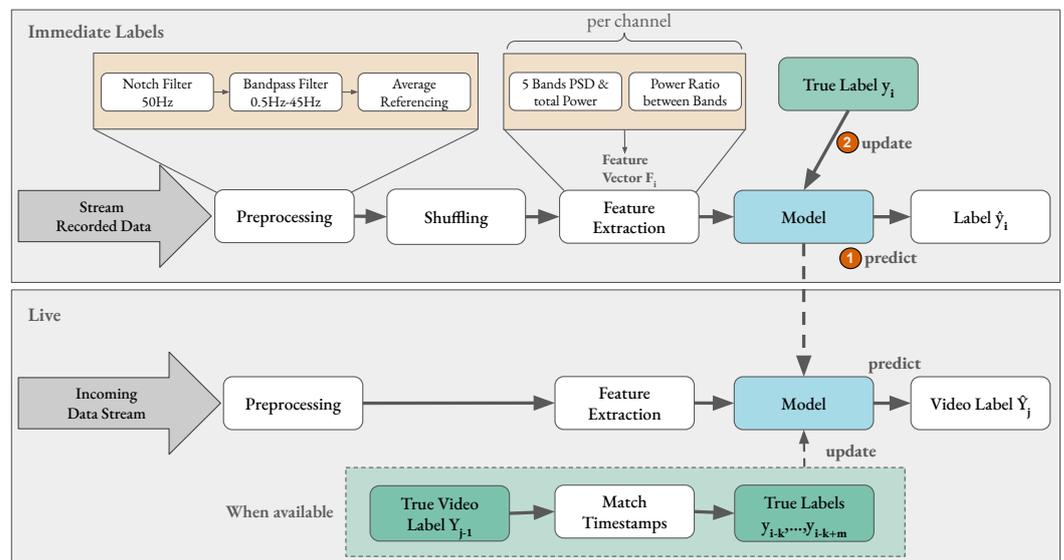
### 2.7. Dataset III: Live Training and Classification

The experimental setup for curating Dataset III is depicted in Figure 5. The participants received the same briefing as mentioned in Section 2.6. For both parts of the experiment, the same device was used. The protocol for the stimuli presentation in the first part (before the break) was the same as that as for the experiment with Dataset II, i.e., a relaxation video, PANAS, eye closing, eight video stimuli, and the AS and familiarity questions. One additional instruction after each AS was shown, which includes the original label of the videos. This additional information was given to the participant since the arousal ratings given in Dataset II were very imbalanced. During the break, the recorded EEG data were pre-processed and used to train an initial model in an online way. This initial model training was necessary because the data needed to be shuffled, as explained in Section 3.2. The initial model was continuously trained and updated during the second part of the experiment where a live prediction of affect is performed. The second part of the experiment was conducted similar to the first part: a relaxation video, eye closing, eight video stimuli, and the AS and familiarity questions. However, one additional prediction was performed and was available to the experimenter before the AS label from the participant. Furthermore, the AS label was used to update the model training and the prediction was running in

parallel. Figure 6 displays the initialized model in the bottom gray rectangle that performed live emotion classification on the incoming EEG data stream. However, the prediction results were only displayed to the experimenter to avoid additional bias. Since the objective of this experiment was *live* online learning and classification, the data were coming in an online stream; however, the data were also stored for later evaluation and reproducibility.



**Figure 5.** Experimental setup for curating Dataset III. In the first session, the participants watched a relaxation video at the beginning and eight videos, two of each dimension category wearing one of the two devices. Between the eight videos, they answered AS slider, familiarity with the video, and had seen the actual AS label. In the second session, they watched the same set of videos while the prediction was available to the experimenter before the delayed label arrived.



**Figure 6.** Overview of pipeline steps for affect classification. The top gray rectangle shows the pipeline steps employed in an immediate label setting with prerecorded data. For each extracted feature vector the model (1) first classifies its label before (2) being updated with the true label for that sample. In the live setting, the model is not updated after every prediction, as the true label of a video only becomes available after the stimulus has ended. The timestamp of the video is matched to the samples' timestamps to find all samples that fall into the corresponding time frame and update the model with their true labels (shown in dotted lines).

In this experiment, the same PC from the previous experiment was again used to present the stimuli. Additionally, the AS label was sent to a second machine (MacBook Pro (2019) with a 2.8 GHz Quad-Core (Intel Core i7) and 16 GB). It also received EEG data, and performed data pre-processing, online model training, and live emotion classification.

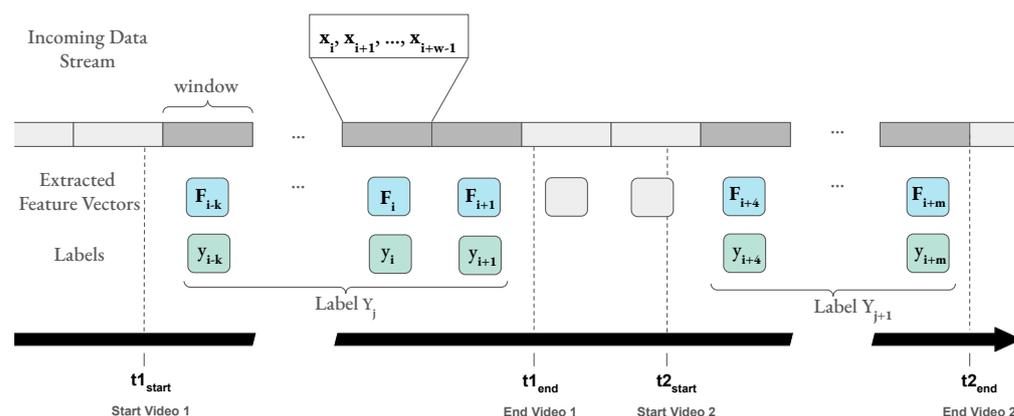
### 3. Emotion Classification Pipeline

#### 3.1. Data Pre-Processing

In this paper, the already pre-processed data from Dataset I were used, whereas Dataset II and Dataset III went through significant pre-processing to remove artifacts from the data [60,61]. Figure 6 depicts the pre-processing steps applied to both the immediate label setting (top) and in a live application (bottom). To remove the power-line interference visible on raw EEG recordings as a sinusoidal at 50 Hz (in Europe) [62], a second-order IIR notch digital filter was applied [63]. Furthermore, a fifth-order Butterworth band pass frequency filter from 0.5 to 45.0 Hz was applied to remove noise on irrelevant frequencies. Additionally, the data were average-referenced after filtering, i.e., the overall average potential was subtracted from each channel [22,37]. This method “relies on the statistical assumption that multichannel EEG recordings are uncorrelated” [64] and assumes an even potential distribution across the scalp. The pre-processing had to be minimal instead of using computation-heavy steps, since the live prediction had to be time efficient.

#### 3.2. Data Windowing and Shuffling

Since EEG data are considered stationary only over short time intervals, the pre-processing and the feature extraction were performed in tumbling windows with a fixed size and no overlap. Figure 7 shows that one window of the incoming data stream includes all samples  $x_i, x_{i+1}, \dots$  arriving during the specified window length. The pipeline extracts one feature vector,  $F_i$ , per window. All feature vectors extracted from the windows of a video duration (between  $t_{start}$  and  $t_{end}$ ) received a label  $y_i$ , corresponding to the reported label,  $Y_j$ , by the participants. Different window lengths,  $l \in [1s, 2s, 3s, 4s, 5s]$ , were tested on Dataset I and Dataset II to find the optimal one for the classification pipeline. As mentioned in the algorithm in Appendix A, a window,  $|w|$ , included  $l \times sf$  samples with the sampling frequency denoted by  $sf$ .



**Figure 7.** The incoming data stream is processed in tumbling windows (gray rectangles). One window includes all samples  $x_i, x_{i+1}, \dots$  arriving during a specified time period, e.g., 1 s. The pipeline extracts one feature vector,  $F_i$ , per window. Windows during a stimulus (video) are marked in dark gray. Participants rated each video with one label per affect dimension,  $Y_j$ . All feature vectors extracted from windows that fall into the time frame of a video (between  $t_{start}$  and  $t_{end}$  of that video) receive a label  $y_i$  corresponding to the reported label,  $Y_j$ , of that video. If possible, the windows are aligned with the end of the stimulus; otherwise, all windows that lie completely inside a video’s time range are considered.

Figure 7 shows that a lot of samples in a row received the same label of the duration of each video up to several minutes. Internal testing implied that training a model by streaming the data resulted in classifiers that did not learn from features but only returned the same class until seeing a different one. Therefore, the time windows were shuffled among one another with the corresponding labels. Since shuffling needs all data and labels to be present before feature extraction, it was not performed during live training and classification.

### 3.3. Feature Extraction

Similar to [22], power spectral density (PSD) features per channel were derived from the raw EEG data by using the Welch method [65] on each window. The PSD from each of the five frequency bands (Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–16 Hz), Beta (16–32 Hz), and Gamma (32–45 Hz)) and the total power over all frequency bands were extracted. Moreover, the power ratios between each pair of frequency bands were obtained. Therefore, a total of sixteen power-related features (five frequency bands + one total power + ten power ratios) were extracted from each channel, resulting in a different number of features per device as depicted in Table 2.

**Table 2.** Number of channels and derived features for each device: Muse Headband: 64 features; Neurosity Crown: 128 features; Emotiv EPOC: 224 features.

Device	Number of Channels	Number of Derived Features
Muse Headband	4	64
Neurosity Crown	8	128
Emotiv EPOC	14	224

### 3.4. Labeling

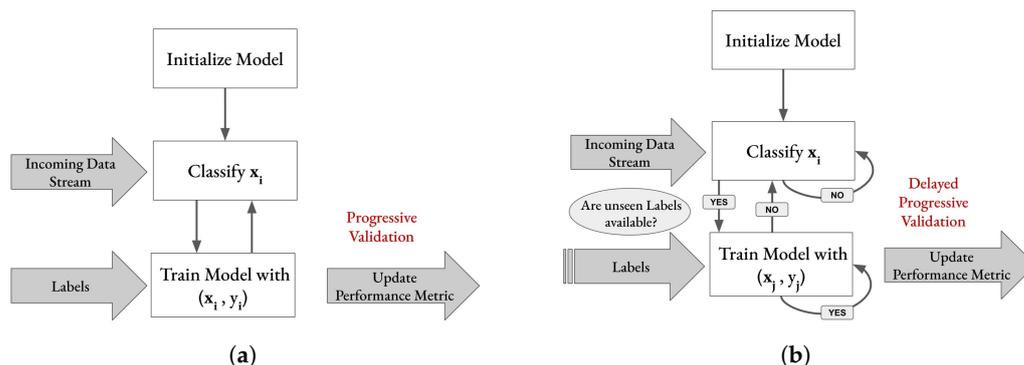
During the live streaming of Dataset III, labels had to be mapped to their corresponding sample. Therefore, the labels were sent in a stream of tuples:  $\mathcal{L}_{1,A}, \mathcal{L}_{1,V}, \mathcal{L}_{2,A}, \mathcal{L}_{2,V}, \dots$ , where

$$\mathcal{L}_{j,dimension} = (Y_{j,dimension}, t_{start}, t_{end}) \quad (1)$$

A and V stand for arousal and valence, respectively, and  $Y_{j,dimension}$  represents the AS label given by the participant after each video of two timestamps,  $t_{start}$  and  $t_{end}$ . One labeled tuple  $\mathcal{L}_{j,dimension}$  per video and dimension was sent from the PC running the PsychoPy experiment to the PC training the classification model. The included timestamps were used to match the incoming ratings,  $Y_{j,dimension}$ , as labels to the samples that the model had classified before. This was performed in a way that all the samples that fell into the time period between  $t_{start}$  and  $t_{end}$  received the respective class label for each dimension. The model could then be updated with these labels.

### 3.5. Evaluation

**Online Learning and Progressive Validation:** This paper aims at building a classification pipeline from evolving data streams. Therefore, the static data from Dataset I and Dataset II were streamed using a library for online learning: *river* [66]. *Progressive validation*, also called *test-then-train evaluation* [67], was used for model evaluation in the supervised immediate label setting, when the labels for all samples were present at processing time [68]. Figure 8a shows the training process of an online classifier including progressive validation. Every time the model sees a new sample  $x_i$ , it first classifies this sample as the test step of the test-then-train procedure. In the training process, the model will calculate the loss by comparing the true label,  $y_i$ , which might come from a different data source than the samples. The updated model will go on to classify the next incoming sample,  $x_{i+1}$ , before seeing its label,  $y_{i+1}$ , and, again, execute the training and performance metric updating step. This continues as long as data are streamed to the model. In this way, all samples can be used for training as well as for validation without corrupting the performance evaluation.



**Figure 8.** (a) Progressive validation incorporated into the basic flow of the training process (“test-then-train”) of an online classifier in an immediate label setting.  $(x_i, y_i)$  represents an input feature vector and its corresponding label. (b) Evaluation incorporated into the basic flow of the training process of an online classifier when labels arrive delayed ( $i \geq j$ ).

In the experimental setup acquiring Dataset III, the labels were available after the prediction, in contrast to the immediate labeling setting and progressive validation. Therefore, a *delayed progressive validation* was performed with the delayed labels, which is mostly the case for real-life scenarios. Figure 8b depicts the delayed progressive validation procedure, where the samples are classified by the model until unseen labels are available. However, the model can be updated as in the immediate label setting. Whenever new labels become available, the performance metric is updated without any further calculations [69]. Once the model has been updated with all available labels, the classification of further samples continues with the new updated model. This can, of course, be implemented in parallel as well. These steps continue as long as there are incoming data.

### 3.6. Machine Learning Classifiers

In this paper, three different algorithms, Adaptive Random Forest (ARF) [26], Streaming Random Patches (SRP) [70], and Logistic Regression (LR), have been evaluated and compared on Dataset I and Dataset II to find the best-performing setup for Dataset III. The ARF and the SRP with a Hoeffding Adaptive Tree (HAT) [42] are two ensemble architectures with integrated drift detection algorithms. Ensemble learners, which combine multiple weak learners, are popular in online learning not only because they tend to achieve high accuracy rates, but also because the individual learners of the ensemble can be trained in parallel. Furthermore, the structure of ensemble learners innately supports drift adaption, as drift detection algorithms can be easily incorporated and component learners can be reset [70,71]. The LR was included in the comparison as a sort of naïve baseline model by training on mini-batches (with partial fit) of one sample (i.e., a feature vector extracted from one window) to resemble the online learning process. Furthermore, it uses stochastic gradient descent for optimization with a learning rate of 0.1; no regularization was applied. For all models, the implementations from the river library [66] were used with default parameters if not specified otherwise.

**Evaluation Metrics:** The participants’ self-reported assessment of their valence and arousal levels was used as the ground truth in all training and evaluation processes in this paper. Among the different metrics of reporting the classifier’s performance [20], the commonly reported metrics Accuracy and F1-Score will be disclosed in this work. They are defined as follows [72]:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$\text{F1-Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}, \quad (3)$$

where TP and TN denote the number of true positives and true negatives classified by the model, respectively. Accordingly, FP and FN stand for the number of false positives and false negatives classified by the model, respectively. “The higher, the better”, can be said for both accuracy and F1-Score, i.e., a perfect model has an accuracy of one (100%) and an F1-Score of one.

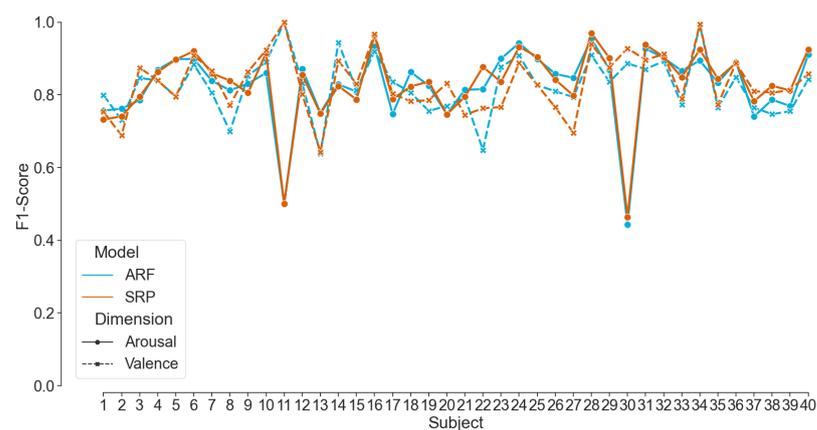
To determine whether the performance differences between the different setups were significant, two-sided t-tests with a significance level of  $\alpha = 0.05$  were conducted on the respective dataset. When important, the results of these tests will be reported by either a  $p > 0.05$ , meaning that no significant differences could be determined at this significance level, or by a  $p < 0.05$ , denoting that the test showed the results of the two compared groups to be significantly different under this test setup.

## 4. Results

### 4.1. Immediate Label Setting

In this paper, at first the real-time emotion classification pipeline with immediate label setting was applied to Dataset I and Dataset II. The data were streamed to pre-process and to extract features from tumbling windows with a window length of 1 s. To perform binary classification for both dimensions of AS (valence and arousal), the self-rating of the participant was used by applying a threshold at 0.5 and defining *high* and *low* classes. As mentioned earlier, ARF, SRP, and LR classifiers were employed for evaluation. The setting of five trees and four trees for SRP worked best for Dataset I and for Dataset II, respectively. ARF included five trees for both data sets. A subject-dependent model was trained with 10-fold cross-validation, and the performances were evaluated with progressive validation.

Figure 9 presents the overall model performances of Dataset I by showing the subject-wise distribution of the evaluation matrix. The mean F1-Score for the positive and negative classes of valence and arousal recognition, respectively, are shown only for ARF and SRP classifiers. The LR showed a comparatively poor performance since it is not an online model but trained on mini-batches. On the contrary, the ensemble models (ARF and SRP) show consistently higher F1-Scores, mostly between 0.7 and 0.95, with two outliers, which validates the emotion classification pipeline built in this paper. Two outliers are visible from subjects 11 and 30 and might be due to a label imbalance (high/low) in the data or insufficient data quality.



**Figure 9.** F1-Score for Valence and Arousal classification achieved by ARF and SRP per subject from Dataset I.

Furthermore, the means of F1-Score and accuracy over all the subjects from Dataset I are presented in Table 3. As depicted in “gray”, both evaluation matrices reach more than 80% for both the ensemble models (ARF and SRP), whereas the performance of LR is relatively poor. Additionally, Table 3 also shows the comparison to the evaluation of the baseline results by Miranda-Correa et al. [22] with a reported approximately 50% F1-Score

with no accuracy score reported. Siddharth et al. [73] reports a more than 70% accuracy and F1-Score, and Topic et al. [74] achieved the current benchmark for this dataset by reporting a 90% accuracy. However, all the related work mentioned was obtained using a hold-out or k-fold cross-validation with computation-heavy offline classifiers, making it inadequate for real-time classification. Consequently, the proposed online classifiers with ensemble models can contribute to daily life setups.

**Table 3.** Comparison of mean F1-Scores and accuracy of Valence and Arousal recognition on Dataset I. Gray represents the results from this paper. NR stands for not reported.

Study or Classifier	F1-Score		Accuracy	
	Valence	Arousal	Valence	Arousal
LR	0.669	0.65	0.702	0.688
ARF	0.825	0.826	0.82	0.846
SRP	0.834	0.831	0.826	0.847
Miranda-Correa et al. [22]	0.576	0.592	NR	NR
Siddharth et al. [73]	0.8	0.74	0.83	0.791
Topic et al. [74]	NR	NR	0.874	0.905

Afterwards, we evaluated the pipeline using Dataset II curated by imitating a real-world setup. Table 4 presents the F1-Score of the subject-dependent models with the three classifiers for the positive and negative classes of arousal and valence recognition, respectively. The F1-Scores from both employed EEG devices are shown, with the best highlighted in bold. As depicted, all classifiers achieved higher performances on arousal recognition than on valence, which is in line with the literature [20,22]. Furthermore, the two ensemble methods ARF and SRP showed a better performance, with a mean F1-Score of more than 82% with no statistically significant difference ( $p > 0.05$ ) between them. LR models showed poor performance similar to their performance on Dataset I. Moreover, the mean F1-Score over all subject-dependent models using Crown data led to a better performance (by at least 2% and up to 7.6%) in most cases compared to using Muse data. However, the differences were not statistically significant ( $p > 0.05$ ) because four out of eleven cases for valence and five out of eleven cases for arousal recognition showed the F1-Score for this dataset. Thus, the distribution of which device's data lead to the best performance is relatively balanced, leading to potential future work of analyzing specific electrode positions.

**Table 4.** Comparison of mean F1-Scores of Arousal and Valence recognition per participant and device from Dataset I with three classifiers using progressive validation. Bold values indicate the best-performing model per participant and dimension. The mean total represents the calculated average of all models' F1-Scores.

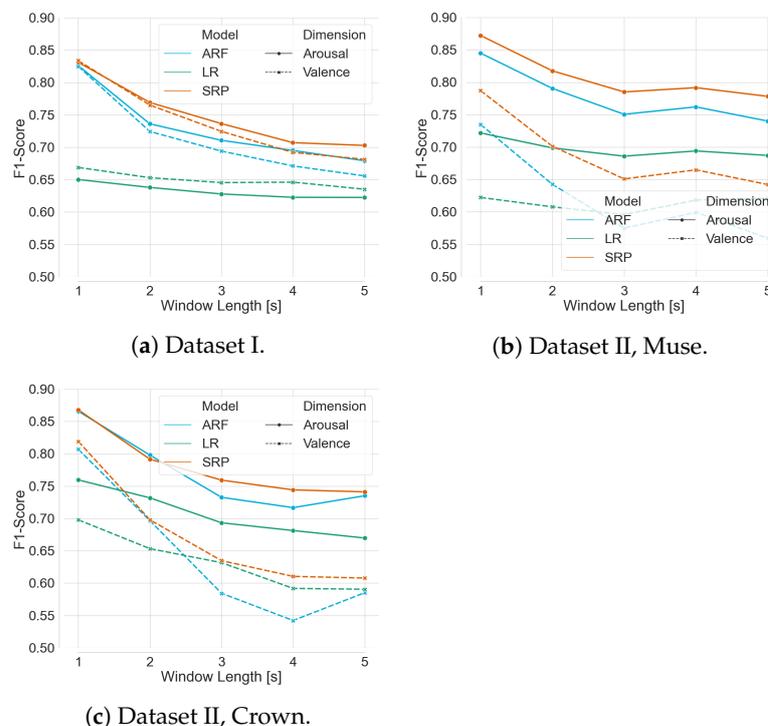
Subject ID	ARF		SRP		LR		
	Crown	Muse	Crown	Muse	Crown	Muse	
Arousal	3	<b>0.902</b>	0.885	0.895	0.898	0.8	0.785
	4	0.836	0.794	0.838	<b>0.845</b>	0.793	0.604
	5	0.651	0.812	0.699	<b>0.827</b>	0.764	0.682
	6	0.836	0.843	0.863	<b>0.889</b>	0.771	0.62
	7	<b>0.958</b>	0.833	0.933	0.878	0.841	0.725
	8	0.889	0.749	<b>0.893</b>	0.783	0.683	0.584
	9	0.888	0.921	0.836	<b>0.931</b>	0.756	0.703
	10	<b>0.969</b>	0.903	0.951	0.915	0.816	0.898
	11	0.938	0.768	<b>0.955</b>	0.861	0.765	0.908
	12	0.864	0.871	<b>0.884</b>	0.878	0.669	0.697
	13	0.792	<b>0.913</b>	0.8	0.887	0.701	0.734
	Mean	0.866	0.845	0.868	<b>0.872</b>	0.76	0.722

Table 4. Cont.

Subject ID	ARF		SRP		LR		
	Crown	Muse	Crown	Muse	Crown	Muse	
Valence	3	0.837	<b>0.887</b>	0.811	0.876	0.716	0.712
	4	0.841	0.69	0.773	<b>0.859</b>	0.804	0.524
	5	0.546	0.734	0.639	0.748	<b>0.781</b>	0.58
	6	0.713	0.687	<b>0.785</b>	0.778	0.73	0.393
	7	<b>0.935</b>	0.666	0.926	0.757	0.776	0.616
	8	0.813	0.551	<b>0.819</b>	0.623	0.594	0.444
	9	0.812	0.844	0.721	<b>0.863</b>	0.72	0.561
	10	<b>0.982</b>	0.859	0.979	0.871	0.74	0.874
	11	0.924	0.653	<b>0.957</b>	0.811	0.64	0.884
	12	0.889	0.756	<b>0.914</b>	0.784	0.633	0.663
	13	0.584	<b>0.826</b>	0.6	0.775	0.543	0.595
	Mean	0.807	0.735	<b>0.819</b>	0.787	0.698	0.622

#### 4.2. Effects of Window Size

As detailed in Section 3, the pipeline processes the incoming data and extracts the features that are used to train the model in tumbling windows of a specified length,  $l \in [1s, 2s, 3s, 4s, 5s]$ . Using Dataset I and Dataset II (both Muse and Crown), the optimum window length was investigated for live prediction. As depicted in Figure 10, with 10-fold cross-validation and progressive validation, the mean F1-Scores from ARF and SRP classifiers show that the best classification performance was achieved with a window length of 1 s irrespective of the affect dimensions, classifiers, and devices. Moreover, in most cases, the classification performance decreases with increasing window size, emphasizing the need for more data points. Furthermore, these plots showcase again that the ensemble methods achieved overall higher F1-Scores than logistic regression and that all classifiers performed better on arousal recognition than on valence.



**Figure 10.** Mean F1-Score achieved by ARF, SRP, and LR over the whole dataset for both affect dimension with respect to window length.

#### 4.3. Delayed Label Setting: Live Classification

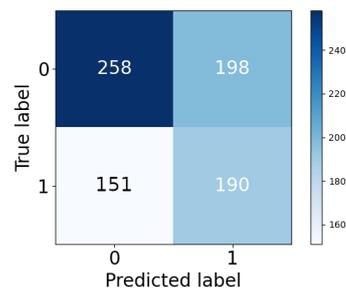
In order to validate the streaming setup of Dataset II, live predictions and live online training were performed while obtaining Dataset III. The participants wore the same EEG device for both parts of the experiment; participants 14 and 17 wore the Muse headband and participants 15 and 16 wore the Crown. For each participant, an ARF with four trees was trained on the data recorded in part 1 of the experiment using a window length of 1 s and progressive delayed validation. With the pre-trained model, live predictions were performed with the data streaming in part 2 of this experiment. The prediction results were only available to the experimenter and the model was continuously updated whenever new true labels became available from the participant. Therefore, the labels arrived with a certain delay depending on the length of the video. Table 5 shows that the highest F1-Score (in bold) obtained from each category during the live prediction was 73% for arousal and 60% for valence. However, most of the reported accuracies in Table 5 barely reached chance level.

**Table 5.** F1-Score and accuracy for the live affect classification (Dataset III, part 2). Subjects 14 and 17 wore Muse, while subjects 15 and 16 wore the Crown for data collection. The highest scores across all participants for each evaluation matrix are marked in bold.

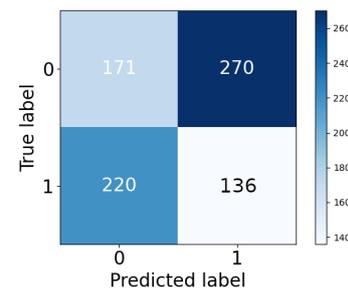
Subject ID	F1-Score		Accuracy	
	Valence	Arousal	Valence	Arousal
14	0.521	0.357	0.562	0.385
15	<b>0.601</b>	0.64	<b>0.609</b>	<b>0.575</b>
16	0.353	<b>0.73</b>	0.502	<b>0.575</b>
17	0.512	0.383	0.533	0.24

Furthermore, the prediction results are presented in the form of confusion matrices in Figure 11, showing that many of the samples have been misclassified. For three out of the four subjects, a low valence could be classified with a recall of at least 0.567, while for the arousal dimension, a low class was misclassified most often with a low recall. Though the proposed prediction pipeline was able to classify valence for over half the samples, it was unsuccessful in reliably predicting accurate classes for both affect dimensions. However, the findings from the live experiment show the importance of testing an application with settings as close as possible to the expected production scenario in order to get a reasonable performance estimate. Moreover, the results display the potential of emotion classification for live applications and motivate us to further investigate frameworks for real use cases with delayed labels instead of solely focusing on the immediate label setting.

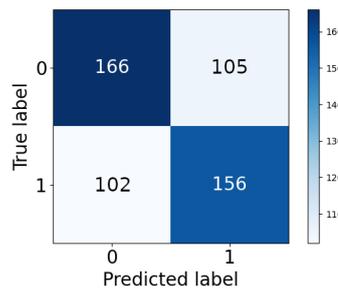
Additionally, the lower predictive performance led us to further investigate the delayed labels. To imitate production settings, we induced a delay into the pipeline and applied progressive delayed validation on Dataset II. Therefore, the subject-dependent model was updated with the true label after 86 samples, i.e., the mean length of the video stimuli was 86 s. Table 6 displays the F1-Scores of both the models for valence and arousal recognition with a label delay of 86 s using an ARF with four trees and a window length of 1 s. The F1-Score for individual participants reached 77% for valence and 78% for arousal. However, the mean F1-Score across all participants achieved 63% for arousal and did not reach chance level for the valence classification. The performance declines significantly compared to Table 4, when a delay is induced.



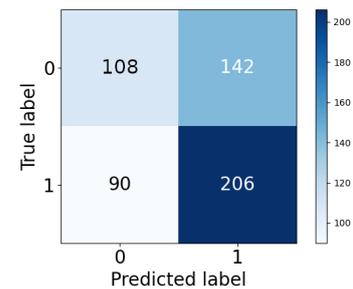
Subject 14, Valence. recall: 0.567



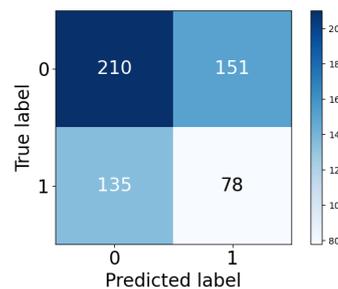
Subject 14, Arousal. recall: 0.387



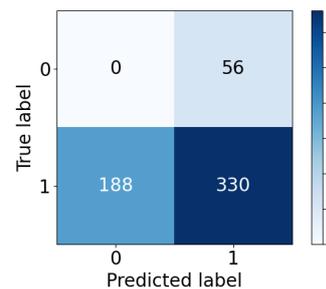
Subject 15, Valence. recall: 0.612



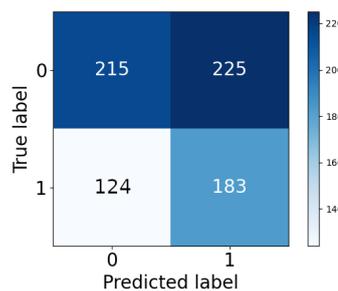
Subject 15, Arousal. recall: 0.432



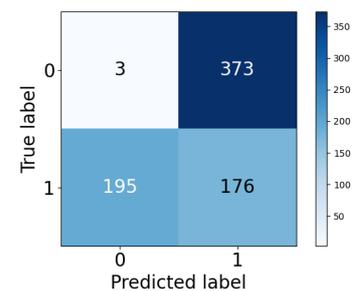
Subject 16, Valence. recall: 0.582



Subject 16, Arousal. recall: 0.0



Subject 17, Valence. recall: 0.489



Subject 17, Arousal. recall: 0.008

**Figure 11.** Confusion matrices for the live affect classification (Dataset III, part 2). Employed model: ARF (four trees), window length = 1 s. Recall was calculated only for a low class for both the models.

However, the findings justify the poor performance in the live settings and validate the pipeline as a useful one with the possibility of modifications in future work. Furthermore, the binary arousal classification with the induced label delay outperforms the baseline results obtained by Miranda-Correa et al. [22] by 4.5% with an immediate label setting. However, the results reported by Siddharth et al. [73] and Topic et al. [74] outperform our work when used with immediate labels.

**Table 6.** Mean F1-Scores for Valence and Arousal recognition of Dataset II, relayed per participant and device. Obtained using ARF (with four trees), a window length of 1 s, and progressive delayed validation with a label delay of 86 s. The bold values represent the highest scores, and the last row shows the mean F1-Score of all participants.

Participant ID	Valence		Arousal	
	Crown	Muse	Crown	Muse
3	0.338	0.584	0.614	0.718
4	0.674	0.429	0.551	0.575
5	0.282	0.554	0.355	0.69
6	0.357	0.27	0.608	0.619
7	0.568	0.574	0.698	0.769
8	0.266	0.286	0.561	0.574
9	0.553	0.53	0.719	0.749
10	<b>0.767</b>	0.561	<b>0.784</b>	0.691
11	0.469	0.207	0.676	0.418
12	0.443	0.51	0.575	0.679
13	0.335	0.451	0.646	0.711
Mean	0.476	0.46	0.637	0.637

## 5. Discussion

In this paper, firstly, a real-time emotion classification pipeline was built for binary classification (high/low) of the two affect dimensions: *Valence* and *Arousal*. Adaptive random forest (ARF), streaming random patches (SRP), and logistic regression (LR) classifiers with 10-fold cross-validation were applied to the EEG data stream. The subject-dependent models were evaluated with progressive and delayed validation, respectively, when immediate and delayed labels were available. The pipeline was validated on existing data of ensured quality from the state-of-the-art AMIGOS [22] dataset. By streaming the recorded data to the pipeline, the mean F1-Scores were more than 80% for both ARF and SRP models. The results outperform the authors' baseline results by approximately 25% and are also slightly better than the work reported in [73] using the same dataset. The results of Topic et al. [74] showed a better performance; however, due to the reported complex setup and computationally expensive methods, the system is unsuitable for real-time emotion classification. Nevertheless, the results mentioned in the related work apply offline classifiers with a hold-out or a k-fold cross-validation technique. In contrast, our pipeline applies an online classifier by employing progressive validation. To the best of our knowledge, no other work has tested and outperformed our online EEG-based emotion classification framework on the published AMIGOS dataset.

Secondly, a similar framework to the AMIGOS dataset, Dataset II, was established within this paper, which can collect neurophysiological data from a wide range of neurophysiological sensors. In this paper, two consumer-grade EEG devices were used to collect data from 15 participants while watching 16 emotional videos. The framework available in the mentioned repository can be adapted for similar experiments.

Thirdly, and most importantly, we curated data in two experiments to validate our classification pipeline using the mentioned framework. Eleven participants took part in acquiring the data for Dataset II, where EEG data were recorded while watching 16 emotion elicitation videos. The pre-recorded data were streamed to the pipeline and showed a mean F1-Score of more than 82% with ARF and SRP classifiers using progressive validation. This finding validates the competence of the pipeline on the challenging dataset from consumer-grade EEG devices. Additionally, the online classifiers consistently showed better performance for ARF and SRP than LR on all compared modalities. However, internal testing verified that the run-time of the training step of the pipeline of ARF is less than that of SRP, concluding that ARF should be used in live prediction. The analysis on window length shows a clear trend of increasing performance scores with decreasing window length; therefore, a window length of 1 s was chosen for live prediction. Although the two

employed consumer-grade devices have a different number of sensors at contrasting positions, there were no statistically significant differences between the achieved performance scores found for their respective data. Therefore, we used both devices for live prediction, and the pipeline was applied to a live incoming data stream in the experiment of Dataset III with the above-mentioned features of the model. In the first part of the experiment, the model was trained with the immediate labels from the EEG data stream. In the second part, the model was used to predict affect dimensions while the labels were available after a delay of the video length. The model was continuously updated whenever a new label is available. The performance scores achieved during the live classification with delayed labels were much lower than those with immediate labels, motivating us to induce an artificial delay to stream Dataset II. The results are compatible with live prediction. The literature reports better results for real-time emotion classification frameworks [29,30,36] with the assumption of knowing the true label immediately after a prediction. The novelty of this paper is to present a real-time emotion classification pipeline close to a realistic production scenario from daily life with the possibility of including further modifications in future work.

In future work, the selected stimuli can be shortened to reduce the delay of the incoming labels so that the model can be updated more frequently. Otherwise, multiple intermediate labels can also be included in the study design to ensure the inclusion of short-term emotions felt while watching the movies. Furthermore, more dynamic pre-processing of the data can be included with feature selection algorithms for better classification and prediction in live settings. Moreover, the collected data from the experiments reveal a strong class imbalance in the self-reported affect ratings for arousal, with high arousal ratings making up 82.96% of all ratings in that dimension. This general trend towards higher arousal ratings is also visible in Dataset I, albeit not as intensely (62.5% high arousal ratings). In contrast, Betella et al. [59] found “a general desensitization towards highly arousing content” in participants. The underrepresented class can be up-sampled in the model training in the future, or basic emotions can be classified instead of arousal and valence dimensions, solving the multi-class problem [75,76]. By including more participants in the future for live prediction, the prediction can be visible to the participant as well to include neuro-feedback. It will also be interesting to see if the predictive performance improves by utilizing additional modalities other than EEG, for example, heart rate and electrodermal activity [20,23,37].

**Author Contributions:** Conceptualization, S.M., F.E.F.S. and B.A.; data curation, S.M. and F.E.F.S.; formal analysis, F.E.F.S.; resources, S.M.; visualization, S.M. and F.E.F.S.; writing—original draft preparation, S.M.; writing—review and editing, S.M., F.E.F.S. and B.A.; supervision, B.A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was (partially) funded by the Hasso-Plattner Institute Research School on Data Science and Engineering. The publication costs were covered by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project number 491466077.

**Institutional Review Board Statement:** The study was conducted in accordance with the Ethics Committee of University of Potsdam (44/2022).

**Informed Consent Statement:** Informed consent was obtained from all participants involved in the study.

**Data Availability Statement:** The dataset is available from the authors upon request for scientific purposes at <https://doi.org/10.5281/zenodo.7398263> (accessed on 18 February 2023). The source code used for analysis in this study can be found at <https://github.com/HPI-CH/EEGEMO> (accessed on 18 February 2023).

**Acknowledgments:** We appreciate the contribution of all the participants who took part in the study for the sake of science. We also acknowledge the researchers of the AMIGOS dataset for making it available to the researchers.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
ARF	Adaptive Random Forest
AS	Affective Slider
EEG	Electroencephalography
HCI	Human–Computer Interaction
HVLA	High Valence Low Arousal (different combinations are possible)
LR	Logistic Regression
OSC	Open Sound Control
PANAS	Positive Furthermore, Negative Affect Schedules
PSD	Power Spectral Density
SRP	Streaming Random Patches

## Appendix A

---

### Algorithm A1: Live Emotion Classification from an EEG Stream.

---

**Input:** Unlabelled EEG data stream  $\mathcal{S} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_i, \dots\}$   
 Stream of true class labels including corresponding stimulus start- and end-times  $\mathcal{L} = \{(\mathcal{Y}_j, \text{startTime}_j, \text{endTime}_j), (\mathcal{Y}_{j+1}, \dots), \dots\}$   
 Sampling frequency  $\text{sf}$   
 Window length  $|w|$   
 Optional: model

**Output:** Predicted binary affect class (valence: 0/1, arousal: 0/1) per window

```

predictions ← Dict();
extractedData ← Dict();
window ← emptyWindow();
windowSize ← sf × |w|;
windowCounter ← 0;
if no model exists
  | model ← initialise-model();
while Stream S has next tuple x do
  timestamp ← current-time();
  window.add(x);
  windowCounter += 1;
  if windowCounter == windowSize
    | preprocess(window);
    | features ← extract-features(window);
    | predictedClass ← predict-one(model, features);
    | display(predictedClass);
    | predictions[timestamp] ← predictedClass;
    | extractedData[timestamp] ← features;
    | windowCounter ← 0;
    | window ← emptyWindow();
if unseen labels available
  | foreach unseen label tuple (Y, startTime, endTime) do
    | | matchedWindows ← match-timesteps(startTime, endTime,
    | | | extractedData);
    | | matchedPredictions ← match-timesteps(startTime, endTime,
    | | | predictions);
    | | for index in length(matchedWindows) do
    | | | performance-metric-update(Y, matchedPredictions[index]);
    | | | train-one(model, Y, matchedWindows[index]);

```

---

## References

1. Picard, R.W. *Affective Computing*; The MIT Press: Cambridge, MA, USA, 2000. [[CrossRef](#)]
2. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80. [[CrossRef](#)]
3. Haut, S.R.; Hall, C.B.; Borkowski, T.; Tennen, H.; Lipton, R.B. Clinical features of the pre-ictal state: Mood changes and premonitory symptoms. *Epilepsy Behav.* **2012**, *23*, 415–421. [[CrossRef](#)] [[PubMed](#)]
4. Kocielnik, R.; Sidorova, N.; Maggi, F.M.; Ouwerkerk, M.; Westerink, J.H.D.M. Smart technologies for long-term stress monitoring at work. In Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems (CBMS), Porto, Portugal, 20–22 June 2013; pp. 53–58. [[CrossRef](#)]
5. Schulze-Bonhage, A.; Kurth, C.; Carius, A.; Steinhoff, B.J.; Mayer, T. Seizure anticipation by patients with focal and generalized epilepsy: A multicentre assessment of premonitory symptoms. *Epilepsy Res.* **2006**, *70*, 83–88. [[CrossRef](#)]
6. Privitera, M.; Haut, S.R.; Lipton, R.B.; McGinley, J.S.; Cornes, S. Seizure self-prediction in a randomized controlled trial of stress management. *Neurology* **2019**, *93*, e2021–e2031. [[CrossRef](#)] [[PubMed](#)]
7. Kotwas, I.; McGonigal, A.; Trebuchon, A.; Bastien-Toniazzo, M.; Nagai, Y.; Bartolomei, F.; Micoulaud-Franchi, J.A. Self-control of epileptic seizures by nonpharmacological strategies. *Epilepsy Behav.* **2016**, *55*, 157–164. [[CrossRef](#)] [[PubMed](#)]
8. Scaramelli, A.; Braga, P.; Avellanal, A.; Bogacz, A.; Camejo, C.; Rega, I.; Messano, T.; Arciere, B. Prodromal symptoms in epileptic patients: Clinical characterization of the pre-ictal phase. *Seizure* **2009**, *18*, 246–250. [[CrossRef](#)]
9. Moontaha, S.; Steckhan, N.; Kappattanavar, A.; Surges, R.; Arnrich, B. Self-prediction of seizures in drug resistance epilepsy using digital phenotyping: A concept study. In Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare, PervasiveHealth '20, Atlanta, GA, USA, 18–20 May 2020; Association for Computing Machinery: New York, NY, USA, 2021; pp. 384–387. [[CrossRef](#)]
10. Levenson, R.; Lwi, S.; Brown, C.; Ford, B.; Otero, M.; Verstaen, A. Emotion. In *Handbook of Psychophysiology*, 4th ed.; Cambridge University Press: Cambridge, UK, 2016; pp. 444–464. [[CrossRef](#)]
11. Liu, H.; Zhang, Y.; Li, Y.; Kong, X. Review on Emotion Recognition Based on Electroencephalography. *Front. Comput. Neurosci.* **2021**, *15*, 758212. [[CrossRef](#)] [[PubMed](#)]
12. Krigolson, O.E.; Williams, C.C.; Norton, A.; Hassall, C.D.; Colino, F.L. Choosing MUSE: Validation of a Low-Cost, Portable EEG System for ERP Research. *Front. Neurosci.* **2017**, *11*, 109. [[CrossRef](#)]
13. Bird, J.J.; Manso, L.J.; Ribeiro, E.P.; Ekárt, A.; Faria, D.R. A Study on Mental State Classification using EEG-based Brain-Machine Interface. In Proceedings of the 2018 International Conference on Intelligent Systems (IS), Funchal, Portugal, 25–27 September 2018; pp. 795–800. [[CrossRef](#)]
14. Teo, J.; Chia, J.T. Deep Neural Classifiers For Eeg-Based Emotion Recognition In Immersive Environments. In Proceedings of the 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE), Shah Alam, Malaysia, 11–12 July 2018; pp. 1–6. [[CrossRef](#)]
15. Gonzalez, H.A.; Yoo, J.; Elfadel, I.M. EEG-based Emotion Detection Using Unsupervised Transfer Learning. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 694–697. [[CrossRef](#)]
16. Hasnul, M.A.; Aziz, N.A.A.; Alelyani, S.; Mohana, M.; Aziz, A.A. Electrocardiogram-Based Emotion Recognition Systems and Their Applications in Healthcare—A Review. *Sensors* **2021**, *21*, 5015. [[CrossRef](#)]
17. Huang, X.; Kortelainen, J.; Zhao, G.; Li, X.; Moilanen, A.; Seppänen, T.; Pietikäinen, M. Multi-modal emotion analysis from facial expressions and electroencephalogram. *Comput. Vis. Image Underst.* **2016**, *147*, 114–124. [[CrossRef](#)]
18. Li, J.; Qiu, S.; Shen, Y.Y.; Liu, C.L.; He, H. Multisource Transfer Learning for Cross-Subject EEG Emotion Recognition. *IEEE Trans. Cybern.* **2019**, *50*, 3281–3293. [[CrossRef](#)] [[PubMed](#)]
19. Hasan, T.F.; Tatum, W.O. Ambulatory EEG Usefulness in Epilepsy Management. *J. Clin. Neurophysiol.* **2021**, *38*, 101–111. [[CrossRef](#)] [[PubMed](#)]
20. Bota, P.J.; Wang, C.; Fred, A.L.N.; Plácido Da Silva, H. A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals. *IEEE Access* **2019**, *7*, 140990–141020. [[CrossRef](#)]
21. Horvat, M.; Dobrinic, M.; Novosel, M.; Jercic, P. Assessing emotional responses induced in virtual reality using a consumer EEG headset: A preliminary report. In Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 21–25 May 2018; pp. 1006–1010. [[CrossRef](#)]
22. Miranda-Correa, J.A.; Abadi, M.K.; Sebe, N.; Patras, I. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. *IEEE Trans. Affect. Comput.* **2021**, *12*, 479–493. [[CrossRef](#)]
23. Laureanti, R.; Bilucaglia, M.; Zito, M.; Circi, R.; Fici, A.; Rivetti, F.; Valesi, R.; Oldrini, C.; Mainardi, L.T.; Russo, V. Emotion assessment using Machine Learning and low-cost wearable devices. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; pp. 576–579. [[CrossRef](#)]
24. Dadebayev, D.; Goh, W.W.; Tan, E.X. EEG-based emotion recognition: Review of commercial EEG devices and machine learning techniques. *J. King Saud-Univ.-Comput. Inf. Sci.* **2022**, *34*, 4385–4401. [[CrossRef](#)]
25. Suhaimi, N.S.; Mountstephens, J.; Teo, J. EEG-Based Emotion Recognition: A State-of-the-Art Review of Current Trends and Opportunities. *Comput. Intell. Neurosci.* **2020**, *2020*, 8875426. [[CrossRef](#)]

26. Gomes, H.M.; Bifet, A.; Read, J.; Barddal, J.P.; Enembreck, F.; Pfahringer, B.; Holmes, G.; Abdesslem, T. Adaptive random forests for evolving data stream classification. *Mach. Learn.* **2017**, *106*, 1469–1495. [[CrossRef](#)]
27. Müller, K.R.; Tangermann, M.; Dornhege, G.; Krauledat, M.; Curio, G.; Blankertz, B. Machine learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring. *J. Neurosci. Methods* **2008**, *167*, 82–90. [[CrossRef](#)]
28. Liu, Y.; Sourina, O.; Nguyen, M.K. Real-Time EEG-Based Human Emotion Recognition and Visualization. In Proceedings of the 2010 International Conference on Cyberworlds CW, Singapore, 20–22 October 2010; pp. 262–269. [[CrossRef](#)]
29. Liu, Y.; Sourina, O. EEG-based subject-dependent emotion recognition algorithm using fractal dimension. In Proceedings of the 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), San Diego, CA, USA, 5–8 October 2014; pp. 3166–3171. [[CrossRef](#)]
30. Lan, Z.; Sourina, O.; Wang, L.; Liu, Y. Real-time EEG-based emotion monitoring using stable features. *Vis. Comput.* **2016**, *32*, 347–358. [[CrossRef](#)]
31. Lan, Z. EEG-Based Emotion Recognition Using Machine Learning Techniques. Ph.D. Thesis, Nanyang Technological University, Singapore, 2018. [[CrossRef](#)]
32. Hou, X.; Liu, Y.; Sourina, O.; Mueller-Wittig, W. CogniMeter: EEG-based Emotion, Mental Workload and Stress Visual Monitoring. In Proceedings of the 2015 International Conference on Cyberworlds (CW), Visby, Sweden, 7–9 October 2015; pp. 153–160. [[CrossRef](#)]
33. Lan, Z.; Liu, Y.; Sourina, O.; Wang, L.; Scherer, R.; Müller-Putz, G. SAFE: An EEG dataset for stable affective feature selection. *Adv. Eng. Inform.* **2020**, *44*, 101047. [[CrossRef](#)]
34. Javaid, M.M.; Yousaf, M.A.; Sheikh, Q.Z.; Awais, M.M.; Saleem, S.; Khalid, M. Real-Time EEG-Based Human Emotion Recognition. In *Neural Information Processing*; Arik, S., Huang, T., Lai, W.K., Liu, Q., Eds.; Springer International Publishing: Cham, Switzerland, 2015; Volume 9492, pp. 182–190. [[CrossRef](#)]
35. Sarno, R.; Munawar, M.N.; Nugraha, B.T. Real-Time Electroencephalography-Based Emotion Recognition System. *Int. Rev. Comput. Softw. IRECOS* **2016**, *11*, 456. [[CrossRef](#)]
36. Bajada, J.; Bonello, F.B. Real-time EEG-based Emotion Recognition using Discrete Wavelet Transforms on Full and Reduced Channel Signals. *arXiv* **2021**, arXiv:2110.05635. [[CrossRef](#)]
37. Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. DEAP: A Database for Emotion Analysis Using Physiological Signals. *IEEE Trans. Affect. Comput.* **2012**, *3*, 18–31. [[CrossRef](#)]
38. Li, J.; Chen, H.; Cai, T. FOIT: Fast Online Instance Transfer for Improved EEG Emotion Recognition. In Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Republic of Korea, 16–19 December 2020; pp. 2618–2625. [[CrossRef](#)]
39. Zheng, W.-L.; Lu, B.-L. Investigating Critical Frequency Bands and Channels for EEG-Based Emotion Recognition with Deep Neural Networks. *IEEE Trans. Auton. Mental Dev.* **2015**, *7*, 162–175. [[CrossRef](#)]
40. Zheng, W.L.; Liu, W.; Lu, Y.; Lu, B.L.; Cichocki, A. EmotionMeter: A Multimodal Framework for Recognizing Human Emotions. *IEEE Trans. Cybern.* **2019**, *49*, 1110–1122. [[CrossRef](#)]
41. Nandi, A.; Xhafa, F.; Subirats, L.; Fort, S. Real-Time Emotion Classification Using EEG Data Stream in E-Learning Contexts. *Sensors* **2021**, *21*, 1589. [[CrossRef](#)]
42. Bifet, A.; Gavaldà, R. Adaptive Learning from Evolving Data Streams. In Proceedings of the 8th International Symposium on Intelligent Data Analysis: Advances in Intelligent Data Analysis VIII, IDA '09, Lyon, France, 31 August–2 September 2009; Springer: Berlin/Heidelberg, Germany, 2009; pp. 249–260. [[CrossRef](#)]
43. Katsigiannis, S.; Ramzan, N. DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals from Wireless Low-cost Off-the-Shelf Devices. *IEEE J. Biomed. Health Inform. JBHI* **2018**, *22*, 98–107. [[CrossRef](#)]
44. Subramanian, R.; Wache, J.; Abadi, M.K.; Vieriu, R.L.; Winkler, S.; Sebe, N. ASCERTAIN: Emotion and Personality Recognition Using Commercial Sensors. *IEEE Trans. Affect. Comput.* **2018**, *9*, 147–160. [[CrossRef](#)]
45. Bradley, M.M.; Lang, P.J. Measuring emotion: The self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* **1994**, *25*, 49–59. [[CrossRef](#)]
46. Watson, D.; Clark, L.A.; Tellegen, A. Development and validation of brief measures of positive and negative affect: The PANAS scales. *J. Personal. Soc. Psychol.* **1988**, *54*, 1063–1070. [[CrossRef](#)]
47. Towle, V.L.; Bolaños, J.; Suarez, D.; Tan, K.B.; Grzeszczuk, R.P.; Levin, D.N.; Cakmur, R.; Frank, S.A.; Spire, J.P. The spatial location of EEG electrodes: Locating the best-fitting sphere relative to cortical anatomy. *Electroencephalogr. Clin. Neurophysiol.* **1993**, *86*, 1–6. [[CrossRef](#)]
48. Peirce, J.; Gray, J.; Simpson, S.; MacAskill, M.; Höchenberger, R.; Sogo, H.; Kastman, E.; Lindeløv, J. PsychoPy2: Experiments in behavior made easy. *Behav. Res. Methods* **2019**, *51*, 195–203. [[CrossRef](#)] [[PubMed](#)]
49. Dan-Glauser, E.S.; Scherer, K.R. The Geneva affective picture database (GAPED): A new 730-picture database focusing on valence and normative significance. *Behav. Res. Methods* **2011**, *43*, 468–477. [[CrossRef](#)] [[PubMed](#)]
50. Kurdi, B.; Lozano, S.; Banaji, M. Introducing the Open Affective Standardized Image Set (OASIS). *Behav. Res. Methods* **2017**, *49*, 457–470. [[CrossRef](#)] [[PubMed](#)]
51. Lang, P.J.; Bradley, M.M.; Cuthbert, B.N. *International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual*; Technical Report; Springer: Berlin/Heidelberg, Germany, 2008. [[CrossRef](#)]
52. Panda, R.; Malheiro, R.; Paiva, R.P. Novel Audio Features for Music Emotion Recognition. *IEEE Trans. Affect. Comput.* **2018**, *11*, 614–626. [[CrossRef](#)]

53. Zhang, K.; Zhang, H.; Li, S.; Yang, C.; Sun, L. The PMemo Dataset for Music Emotion Recognition. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, ICMR '18, Yokohama, Japan, 11–14 June 2018; ACM: New York, NY, USA, 2018; pp. 135–142. [[CrossRef](#)]
54. Abadi, M.K.; Subramanian, R.; Kia, S.M.; Avesani, P.; Patras, I.; Sebe, N. DECAF: MEG-Based Multimodal Database for Decoding Affective Physiological Responses. *IEEE Trans. Affect. Comput.* **2015**, *6*, 209–222. [[CrossRef](#)]
55. Soleymani, M.; Lichtenauer, J.; Pun, T.; Pantic, M. A Multi-Modal Affective Database for Affect Recognition and Implicit Tagging. *Affect. Comput. IEEE Trans.* **2012**, *3*, 42–55. [[CrossRef](#)]
56. Verma, G.; Dhekane, E.G.; Guha, T. Learning Affective Correspondence between Music and Image. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 3975–3979. [[CrossRef](#)]
57. Russell, J.A. A circumplex model of affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161. [[CrossRef](#)]
58. Breyer, B.; Bluemke, M. Deutsche Version der Positive and Negative Affect Schedule PANAS (GESIS Panel). In *Zusammenstellung Sozialwissenschaftlicher Items und Skalen*; Social Science Open Access Repository (SSOAR): Mannheim, Germany, 2016. [[CrossRef](#)]
59. Betella, A.; Verschure, P. The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions. *PLoS ONE* **2016**, *11*, e0148037. [[CrossRef](#)]
60. Jiang, X.; Bian, G.B.; Tian, Z. Removal of Artifacts from EEG Signals: A Review. *Sensors* **2019**, *19*, 987. [[CrossRef](#)]
61. Sörnmo, L.; Laguna, P. Chapter 3—EEG Signal Processing. In *Bioelectrical Signal Processing in Cardiac and Neurological Applications*; Biomedical Engineering; Academic Press: Burlington, NJ, USA, 2005; pp. 55–179. [[CrossRef](#)]
62. Akwei-Sekyere, S. Powerline noise elimination in neural signals via blind source separation and wavelet analysis. *PeerJ PrePrints* **2014**, *3*. [[CrossRef](#)]
63. Sweeney, K.; Ward, T.; Mcloone, S. Artifact Removal in Physiological Signals-Practices and Possibilities. *IEEE Trans. Inf. Technol. Biomed. Publ. IEEE Eng. Med. Biol. Soc.* **2012**, *16*, 488–500. [[CrossRef](#)] [[PubMed](#)]
64. Yao, D.; Qin, Y.; Hu, S.; Dong, I.; Vega, M.; Sosa, P. Which Reference Should We Use for EEG and ERP practice? *Brain Topogr.* **2019**, *32*, 530–549. [[CrossRef](#)] [[PubMed](#)]
65. Welch, P. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electroacoust.* **1967**, *15*, 70–73. [[CrossRef](#)]
66. Montiel, J.; Halford, M.; Mastelini, S.M.; Bolmier, G.; Sourty, R.; Vaysse, R.; Zouitine, A.; Gomes, H.M.; Read, J.; Abdessalem, T.; et al. River: Machine learning for streaming data in Python. *J. Mach. Learn. Res.* **2021**, *22*, 4945–4952. [[CrossRef](#)]
67. Grzenda, M.; Gomes, H.M.; Bifet, A. Delayed labelling evaluation for data streams. *Data Min. Knowl. Discov.* **2020**, *34*, 1237–1266. [[CrossRef](#)]
68. Blum, A.; Kalai, A.T.; Langford, J. Beating the hold-out: Bounds for K-fold and progressive cross-validation. In Proceedings of the Twelfth Annual Conference on Computational Learning Theory COLT '99, Santa Cruz, CA, USA, 7–9 July 1999.
69. McMahan, H.B.; Holt, G.; Sculley, D.; Young, M.; Ebner, D.; Grady, J.; Nie, L.; Phillips, T.; Davydov, E.; Golovin, D.; et al. Ad Click Prediction: A View from the Trenches. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Chicago, IL, USA, 11–14 August 2013; pp. 1222–1230. [[CrossRef](#)]
70. Gomes, H.M.; Read, J.; Bifet, A. Streaming Random Patches for Evolving Data Stream Classification. In Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM), Beijing, China, 8–11 November 2019; pp. 240–249. [[CrossRef](#)]
71. Parker, B.; Khan, L. Detecting and Tracking Concept Class Drift and Emergence in Non-Stationary Fast Data Streams. In Proceedings of the AAAI, Austin, TX, USA, 25–30 January 2015. [[CrossRef](#)]
72. Aggarwal, C.C. *Data Classification: Algorithms and Applications*, 1st ed.; Chapman & Hall/CRC: Boca Raton, FL, USA, 2014; pp. 636–638.
73. Siddharth, S.; Jung, T.P.; Sejnowski, T.J. Utilizing Deep Learning Towards Multi-Modal Bio-Sensing and Vision-Based Affective Computing. *IEEE Trans. Affect. Comput.* **2019**, *13*, 96–107. [[CrossRef](#)]
74. Topic, A.; Russo, M. Emotion recognition based on EEG feature maps through deep learning network. *Eng. Sci. Technol. Int. J. JESTECH* **2021**, *24*, 1442–1454. [[CrossRef](#)]
75. Liu, Y.J.; Yu, M.; Zhao, G.; Song, J.; Ge, Y.; Shi, Y. Real-Time Movie-Induced Discrete Emotion Recognition from EEG Signals. *IEEE Trans. Affect. Comput.* **2018**, *9*, 550–562. [[CrossRef](#)]
76. Ekman, P.; Friesen, W. *Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues*; Prentice-Hall: Oxford, UK, 1975.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.