Article

Online Learning for Wearable EEG-based Emotion Classification

Sidratul Moontaha [1,†*], Franziska Elisabeth Friederike Schumann [1,†] and Bert Arnrich [1,*]

[1] Digital Health - Connected Healthcare, Hasso Plattner Institute, University of Potsdam, 14482 Potsdam, Germany.
* Correspondence: eegemo@hpi.de (F.S. and S.M.); firstname.lastname@hpi.de (S.M.and B.A.)
† These authors contributed equally to this work.

Abstract: Emotions are indicators of affective states and play a significant role in human daily life, behavior, and interactions. Giving emotional intelligence to the machines could, for instance, facilitate early detection and prediction of (mental) diseases and symptoms. Electroencephalography (EEG) -based emotion recognition is being widely applied because it measures electrical correlates directly from the brain rather than the indirect measurement of other physiological responses initiated by the brain. The recent development of non-invasive and portable EEG sensors makes it possible to use them in real-time applications. Therefore, this paper presents a real-time emotion classification pipeline, which trains different binary classifiers for the dimensions of Valence and Arousal from an incoming EEG data stream. After achieving a 23.9% (Arousal) and 25.8% (Valence) higher f1-score on the state-of-art AMIGOS dataset, this pipeline was applied to the dataset achieved by an emotion elicitation experimental framework developed within the scope of this thesis. Following two different protocols, 15 participants were recorded using two different consumer-grade EEG devices while watching 16 short emotional videos in a controlled environment. For an immediate label setting, the mean f1-score of 87% and 82% were achieved for Arousal and Valence, respectively. In a live scenario, while continuously being updated on the incoming data stream with delayed labels, the pipeline proved to be fast enough to achieve predictions in real time. However, the significant discrepancy from the readily available labels on the classification scores leads to future work to include more data with frequent delayed labels in the live settings.

Keywords: Online Learning; Emotion Classification; AMIGOS dataset; Wearable-EEG (Muse and Neurosity Crown); Psychopy Experiments

1. Introduction

Emotions are part of everyone's daily life as they are crucial to many aspects: They are a significant factor in human interactions, influence decision-making, and are involved in mental health. Emotions play a crucial role in human communication and cognition, which makes comprehending them significant to understanding human behaviour [1]. The field of affective computing strives to build systems that can recognize and interpret human affects [1,2], offering exciting possibilities for education, entertainment, and healthcare. Giving machines emotional intelligence could, for instance, facilitate early detection and prediction of (mental) diseases or their symptoms since specific emotional and affective states are often indicators thereof [3]. For example, 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]. The association of premonitory symptoms and seizure counts has been analyzed from patient reported diaries [6], and the non-pharmacological interventions proved to reduce the seizure rate [7]. However, many PWE can not consistently identify their prodromal symptoms, and many do not perceive prodromes [8], emphasizing the necessity of objective prediction of epileptic seizures. In previous work, the authors proposed

developing a system to predict seizures by continuously monitoring their affective states [9]. Therefore, measuring and predicting affective states in real-time through neurophysiological data could aid in finding pre-emptive therapies for epilepsy patients by identifying the pre-ictal state to predict a seizure onset. That would be incredibly beneficial, especially to people with drug-resistant epilepsy, and improve their quality of life by enabling them to anticipate and mitigate possibly violent seizures [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 interaction (HCI) and enhance their abilities 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]. It 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]. 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 like the *Emotiv EPOC* Neuroheadset or the *Muse S* headband have become quite popular tools in emotion recognition [19–21]. *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) [22], could be reliably inferred through the use of a properly trained classifier.

However, a challenging but important step to identifying stress-related events or improving HCI in real-life settings is to recognise changes in peoples' affect by leveraging live data. The EEG-based emotion classification mentioned in the literature has nearly exclusively employed traditional machine learning strategies, i.e., offline classifiers, often combined with complex data preprocessing techniques, on static datasets, making it unsuitable for daily monitoring. Therefore, researchers are interested in building a real-time emotion classification pipeline since lately, where the classification results are obtained from pre-recorded (and already preprocessed) data, often utilizing a pre-trained model [23,24] rather than working with (live) data streams. Li et al. [25] address the challenges when a model can see the data only once by leveraging cross-subject and cross-session data but does not apply live incoming data stream to their work. Whereas Lan et al. [26] analyse stable features for real-time emotion recognition and implement their proposed algorithm in two emotion-monitoring applications where computer avatars reflect a person's emotion based on live EEG data. However, the live emotion classification is based on a static model, which has to be trained in a prior training session and is not updated afterward. To the best of our knowledge, only Nandi et al. [27] 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 did not report on undertaking any such live experiments. Indeed, more research is needed on using online machine learning for emotion recognition.

Moreover, multi-modal labeled data for the prediction of affective states have been made freely available through annotated affective databases, like DEAP [28], DREAMER [29], ASCERTAIN [30], and AMIGOS [21], which play a significant role in further enhancing the research of this field. They include diverse data from experimental setups using differing emotional stimuli like 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 the

 doi:10.20944/preprints202301.0156.v1

datasets have published the data collection framework to be reused in curating the data from wearable EEG devices in live settings.

Therefore, *firstly*, the key contribution of this paper is the establishment of a lightweight emotion classification pipeline that can provide predictions on a person's affective state based on an incoming EEG data stream in real-time, efficiently enough to be used in real applications i.e., 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 the existing datasets of established quality, AMIGOS, with better predictive 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 of the subjects' affective states even when labels arrive with a delay. Data from 15 participants were captured by using two consumer-grade 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 prediction performance with the AMIGOS dataset. Following this, the live 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 is 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). The Python code for loading the dataset and implementations of the developed pipeline are made available on GitHub (https://github.com/HPI-CH/EEGEMO). The next section will explain the material and methods utilized within this paper following the results and discussion sections.

2. Materials and Methods

2.1. AMIGOS dataset

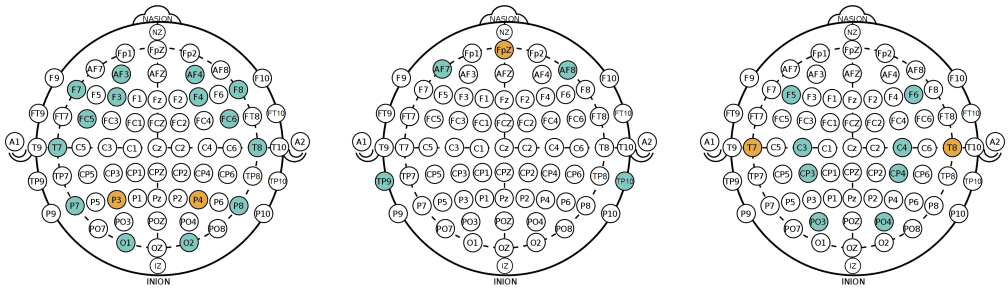
The emotion classification pipeline developed within the scope of this paper was evaluated on the state-of-art dataset for affect, personality, and mood research on individuals and groups (AMIGOS) published by Miranda-Correa et al. [21]. Upon following the data receiving protocol, all data from the AMIGOS dataset that is used in this work stems from the 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 was recorded using the Emotiv EPOC Neuroheadset [1] with a sampling frequency of 128 Hz with 14 bit resolution. This device records EEG data from 14 channels (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4) of the brain according to the 10-20 system as depicted in Figure 1a.

Additionally, AMIGOS dataset reports the Self-Assessment Manikin [31] 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 their liking of them and had to select at least one option from a list of basic emotions that they felt after watching each video. However, only the obtained valence and arousal ratings are considered as the ground truth for the classifier while working with AMIGOS dataset within this paper. Furthermore, the participants answered the Positive and Negative Affect Schedules (PANAS) [32] questionnaire at the beginning of the experiment and a second time in the days after the experiment; only one overall calculated PANAS score is reported. To evaluate the classification pipeline, the preprocessed data files were used where the EEG data was downsampled to 128 Hz, averaged to common reference, and applied a band pass frequency filter from $4.0 - 45.0$ Hz as described in the AMIGOS dataset description website

[1] https://www.emotiv.com/epoc-x/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2023 doi:10.20944/preprints202301.0156.v1

[2]. The files containing electrocardiogram (ECG) and galvanic skin response (GSR) data have been remmoved for the analysis of this paper.



(a) Emotiv EPOC **(b)** Muse S Headband **(c)** Neurosity Crown

Figure 1. Different electrode positions, according to the international 10-20 system, of the EEG devices used in AMIGOS dataset: 1a, and in Experiments I and II: 1b, 1c . Sensor locations are marked in blue, references in orange.

2.2. Experimental Setup

This paper establishes a lightweight emotion classification pipeline that can provide predictions on a person's affective state based on an incoming EEG data stream in real-time efficiently enough to be used in a live setting.Two different experimental protocols, named *Experiment I* and *Experiment II*, were designed with the description of participants, data acquisition, and experimental protocols mentioned below.

2.2.1. Participants

For Experiment I, 13 participants were recruited including two test participants. Therefore, the data analysis cohort consists of 11 participants (6 females and 5 males) between the age of 25 and 42 ($\mu = 29.27$, $\sigma = 5.41$ years). Experiment II was conducted with 4 participants (1 female and 3 males) between the age of 25 and 34 ($\mu = 28.5$, $\sigma = 3.5$ years). Exclusion criteria for the study included being pregnant, being older than 65 years, and had taken part in Experiment II. All participants had normal or corrected vision and reported no history of neurological or mental illnesses or head injuries.

2.2.2. Data Acquisition

Hardware: During the experiments, two consumer-grade devices: *Muse S Headband* Gen 1 [3] and *Neurosity Crown* [4], were used to collect the EEG data from the participants as depicted in Figure 2. While both devices operate with the sampling rate of 256 Hz, the EEG data is collected with 4 and 8 channels, respectively. According to the international 10-20

(a) Muse S Headband **(b)** Neurosity Crown

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

system [33], 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

[2] http://www.eecs.qmul.ac.uk/mmv/datasets/amigos/readme.html
[3] https://choosemuse.com/compare/
[4] https://neurosity.co/crown

 doi:10.20944/preprints202301.0156.v1

Crown are C3, C4, CP3, CP4, F5, F6, PO3, and PO with the reference sensors located at T7 and T8 as shown in Figure 1c. Using Mind Monitor App [5], the raw EEG data was 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 [6] on the receiving end. As the incoming data tuples from the Muse Monitor App did not include timestamps, one was 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 preprocessing settings. In contrast to the Muse Headband, the Crown includes a timestamp when sending data. In order to compare the data from the different devices and have consistent results, the pipeline also added a timestamp to each sample when receiving the data.

Software: In this paper, the experiment was implemented using the software PsychoPy (v 2021.2.3) [34] in a way that guided the participants through instructions, questionnaires, and stimuli. The participants are allowed to go through their own pace by clicking on the 'Next' (in German 'Weiter') button, as shown in the screenshots of PsychoPy in figure Figure 4.

2.2.3. Stimuli Selection

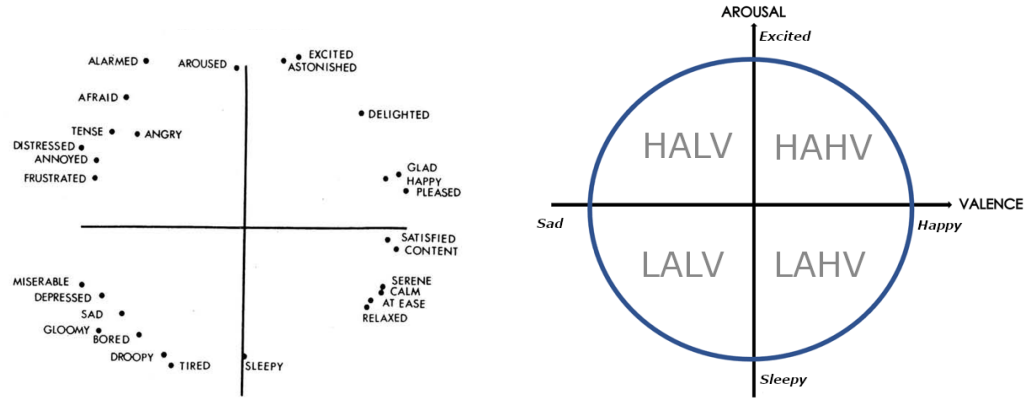
Inducing (specific) emotional reactions, even in a fully controlled experimental setting, is a challenge. Several datasets are trying to solve it with different modalities like pictures [35–37], music [38,39], or (music-) videos [28,40,41] or combinations of them [42]. In this work, videos depicting short movie scenes were used as stimuli, based on the experimental setup Miranda-Correa et al. used for AMIGOS dataset [21]. Therefore, 16 short clips (51-150 s long, $\mu = 86.7$ s, $\sigma = 27.8$ s) depicting scenes from 15 different movies were used in the experiments for emotion elicitation. 12 of these videos stem from the DECAF dataset [40], and 4 movie scenes were taken from the MAHNOB-HCI [41] 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. It means they represent the most extreme ratings in their respective category according to the labels provided by 72 volunteers [21]. 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 [43]. Valence, the dimension describing one's level of pleasure, ranges from sad (unpleasant, stressed) to happy (pleasant, content), and can be seen on the horizontal axis, whereas Arousal, which ranges from sleepy (bored, inactive) to excited (alert, active), is placed on the vertical axis. Experienced affective states can be objectively described by assigning a rating in both dimensions. As depicted in Figure 3, the model divides the plane into four quadrants. Independent of the employed scale, everything greater than the middle of each axis (i.e., above or to the right of it) respectively is usually deemed as a high level of feelings in the corresponding dimension and everything under that threshold is deemed as low. Following this, the four quadrants are called: 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, 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.

2.2.4. Behavioral Data

PANAS: During the experiments, participants were asked to assess their baseline levels of affect, also referred to as mood, in the PANAS scale. As depicted in one of the screenshots from PsychoPy in figure Figure 4a, the total 20 questions (10 question from each of the Positive Affect (PA), and Negative Affect (NA) dimension) were asked to rate in a 5-point Likert scale with the options ranging from "very slightly or not at all" (1) to "extremely"

[5] https://mind-monitor.com/
[6] https://pypi.org/project/python-osc/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2023 doi:10.20944/preprints202301.0156.v1



(a) (b)

Figure 3. Figure 3a Russell's Circumplex Model of Affect in multidimensional scaling [43], 28 affect words are placed on the plane spanned by two axes without explicit title. Figure 3b is a reduced version of Russell's Circumplex Model of Affect [44] depicting the valence-arousal space as it is used in this work with the four quadrants: HALV, HAHV, LAHV, LALV. H, L, A, and V stand for high, low, arousal and valence respectively.

Table 1. The source movies of the videos used in the experiments 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 [41]; all the others stem from DECAF [40]. 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. [21].

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

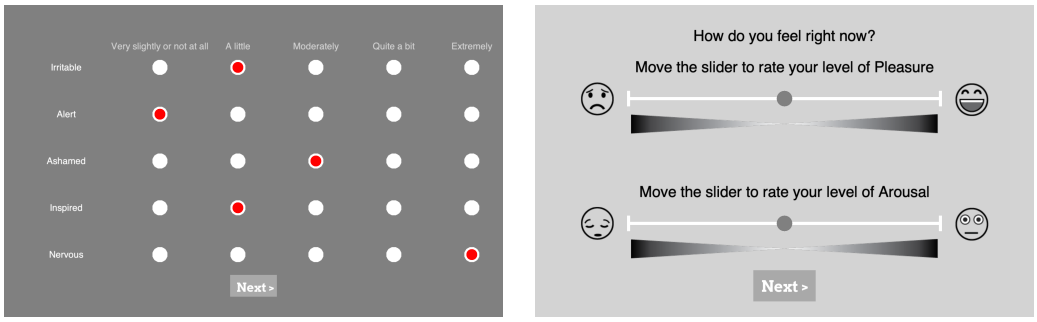
(5). To see if the participants' moods generally changed over the course of the experiments, they were asked to answer the PANAS once at the beginning of the experiment and then once again at the end. For the German version of the PANAS questionnaire, the translation of Breyer and Bluemke [45] was used in the experiments.

Affect Self-Assessment: The Affective Slider (AS) [46] was used in the experiment to capture participants' emotional self-assessment after presenting each stimulus as depicted in the screenshot in Figure 4b [7]. 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 [43]. To rate the experienced level of one of these dimensions, the corresponding slider can be moved to the appropriate point on the scale. For the experiments, AS was designed in a continuous normalised scale with a step size of 0.01 (i.e., a resolution of 100) and the order of the two sliders were randomized each time.

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 used in the experiments and was always shown after the affect self-assessment.

[7] https://github.com/albertobeta/AffectiveSlider

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2023 doi:10.20944/preprints202301.0156.v1



(a) PANAS questionnaire. **(b)** Affective slider.

Figure 4. Screenshots from the PsychoPy [34] experimental setup of self-assessment questions the participants were shown in Experiment I and II. 4a is one part of the PANAS questionnaire with 5 different levels represented by clickable radio buttons with levels explanation on top. 4b shows the AS for valence displayed on top and the slider for arousal on the bottom.

2.2.5. Experiment I

Briefing Session: At the beginning of the experiment, the participant went through a briefing session from the experimenter. In this session, the experimenter explained the study procedure after leading the participant into the study room. The participants were informed that the experiment would entail two parts of approximately 20 minutes each with a small intermediate break. The participant will then receive and read the data information sheet, fill out the personal information sheet, and sign the consent to participate in the study. Personal information includes 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 explains the three scales mentioned in and allows the participant to accustom to the study platform, PsychoPy. This ensures the understanding of the different terms and scales used for the experiment without having to interrupt the experiment afterwards. The participant can refrain from participating at any moment during the experiment.

Data Collection: After briefing, the experimenter put either the Muse Headband or the Neurosity Crown on the participant by a random choice. Putting headphones over the device, the participants was asked to refrain from strong movements, especially with the head. The experimenter then checks the incoming EEG data and let the participant begin with the experiment. After greeting the participant with a welcome screen, a relaxation video were shown to the participant [8].They answered the PANAS questionnaire to rate their current mood and close their eyes for half a minute to get a baseline measure of EEG data. Afterwards, they were asked to initially rate their valence and arousal state with the AS. Following this, an instruction about watching 8 short videos is provided. Each of those was preceded by a video counter and followed by two questionnaires,the AS and the familiarity with each video. 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 label of the videos are balanced. The first part of the experiment ended after watching 8 videos and answering corresponding questionnaire. The participants were allowed a short break after taking the EEG device and the headphone 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 participants followed exact same protocol, watching relaxation video, answering PANAS, closing eyes and watching 8 more movie scenes with the AS and familiarity question in between. Lastly, they were asked

[8] https://www.youtube.com/watch?v=S6jCd2hSVKA

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2023 doi:10.20944/preprints202301.0156.v1

for a final mood self-assessment via a second PANAS questionnaire to capture differences before and after the experiment.

In Experiment I, 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 obtain the subjects' self-assessments through PsychoPy, as well as receive the signals from the EEG measuring devices. All data was stored and only used after the experiment session.

2.2.6. Experiment II : Live Training and Classification

In Experiment II, the participants received the same briefing as mentioned in Section 2.2.5. For both part of the experiment same device has been used in this experiment. The protocol for the stimuli presentation in the first part (before the break) was also followed according to Experiment I: relaxation video, PANAS, eye closing, 8 video stimuli with the AS and familiarity question. 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 Experiment I were very imbalanced. During the break, the recorded EEG data was preprocessed and used to train a initial model in an online way. This initial model training was necessary because the data needed to be shuffled, as explained in Section 2.3.2. The initial model is continuously trained and updated during the second part of the experiment where the live prediction of affect is available. The second part of the experiment is conducted according to the first part: relaxation video, PANAS, eye closing, 8 video stimuli with the AS and familiarity question. However, one additional prediction is performed and 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 5 displays the initialised model in the bottom grey rectangle to do 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 Experiment II was *live* online learning and classification, the data coming in a online stream, however, the data was also stored for later evaluation and reproducibility.

In Experiment II, the same PC that was employed in Experiment I (2, 4 to 3, 0 GHz Dual Intel Core i5-6300U and 12 GB RAM) was used to present the stimuli through PsychoPy, and send the AS label to a second PC. This second machine was a MacBook Pro (2019) with a 2, 8 GHz Quad-Core (Intel Core i7) and 16 GB of memory. It was used to receive the signals from the EEG devices and the labels from the first PC, as well as for data preprocessing, online model training and live emotion classification.

2.3. Emotion Classification Pipeline

2.3.1. Data Preprocessing

In this paper, no additional preprocessing was performed in AMIGOS dataset, since the preprocessed data provided by the author was used. However, the EEG data collected during Experiment I and II went through significant preprocessing to remove artifacts from the data [47,48]. Figure 5 depicts all the similar preprocessing steps applied on both shows the immediate labelling setting (top) and in a live application (bottom). To remove the powerline interference visible on raw EEG recordings as a sinusoidal at 50 Hz (in Europe) [49], a second-order IIR notch digital filter was applied to the data [50]. Furthermore, a fifth-order Butterworth bandpass frequency filter from 0.5 to 45.0 Hz was applied to remove noise on frequencies that were not relevant (see **??**). Additionally, the data was average-referenced after filtering, i.e., the overall average potential is subtracted from each channel [21,28**?**]. This method "relies on the statistical assumption that multichannel EEG recording are uncorrelated" [51] and assumes an even potential distribution across the scalp. The preprocessing had to be minimum instead of computation-heavy steps, since the live prediction had to be time efficient.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2023 doi:10.20944/preprints202301.0156.v1

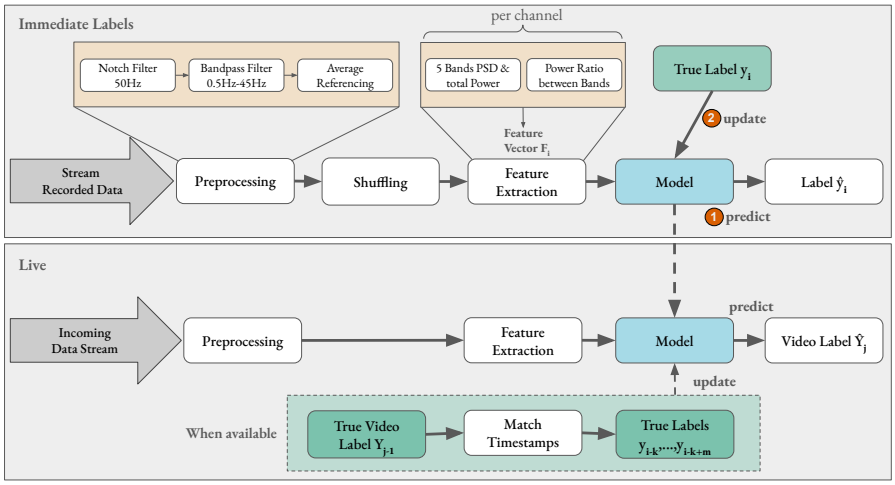


Figure 5. Overview of pipeline steps for affect classification. The top grey rectangle shows the pipeline steps employed in an immediate label setting with prerecorded data. For each extracted feature vector the model (1) first predicts its label before (2) being updated with the true label for that 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 stimuli has ended. The timestamp of the video is matched to the samples' timestamps to find all samples that fell into the corresponding time frame and update the model with their true labels.

2.3.2. Data Windowing and Shuffling

Since EEG data is considered stationary only over short time intervals, the preprocessing and the feature extraction were performed in tumbling windows with a fixed size and no overlap. Figure 6 shows that one window of the incoming data stream includes all samples $\mathbf{x}_i, \mathbf{x}_{i+1}, ...$ arriving during the specified window length. The pipeline extracts one feature vector, $\mathbf{F_i}$, per window. All feature vectors extracted from the windows of a video duration (between t_{start} and t_{end}) receive a label y_i corresponding to the reported label, Y_j by the participants. Different window length, $l \, \epsilon \, [1s, 2s, 3s, 4s, 5s]$ were tested on the AMIGOS dataset and the dataset from Experiment I to find the optimal one for the classification pipeline. As mentioned in the algorithm in Appendix A, a window, $|w|$ includes $l * sf$ samples with the sampling frequency denoted by sf.

Figure 6 shows that a lot of samples in a row received the same label of the duration of each video upto several minutes. Through an internal testing implies that training a model by streaming the data resulted in classifiers that did not learn from features but only always predicted the same value until seeing a different one. Therefore, the time windows are shuffled among one another with the corresponding labels. Since shuffling needs all data and labels present before feature extraction, it was not performed during the live training and classification.

2.3.3. Feature Extraction

Similar to [21], power spectral density (PSD) features per challen were derived from the raw EEG data by using Welch method [52] 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 ratio between each pair of frequency bands was obtained. Therefore, total 16 power related features (5 frequency bands + 1 total power + 10 power ratios) were extracted from each channel resulting different number of features per device as depicted in Table 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2023 doi:10.20944/preprints202301.0156.v1

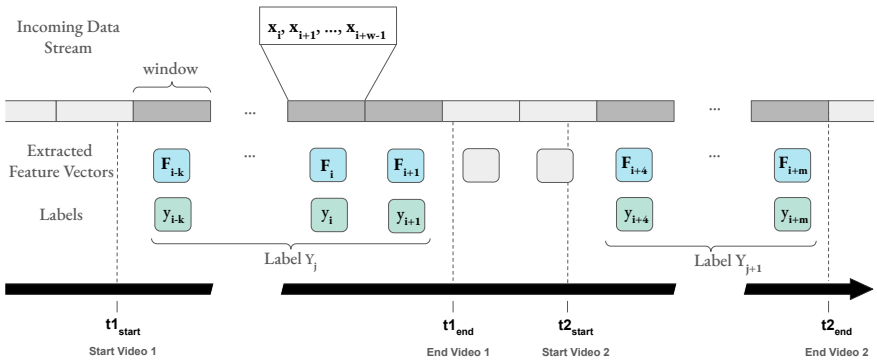


Figure 6. The incoming data stream is processed in tumbling windows (grey rectangles). One window includes all samples $\mathbf{x}_i, \mathbf{x}_{i+1}, ...$ arriving during a specified time period, e.g., 1 second. The pipeline extracts one feature vector $\mathbf{F}_i$ per window. Windows during a stimulus (video) are marked in dark grey. 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.

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

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

2.3.4. Labelling

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

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

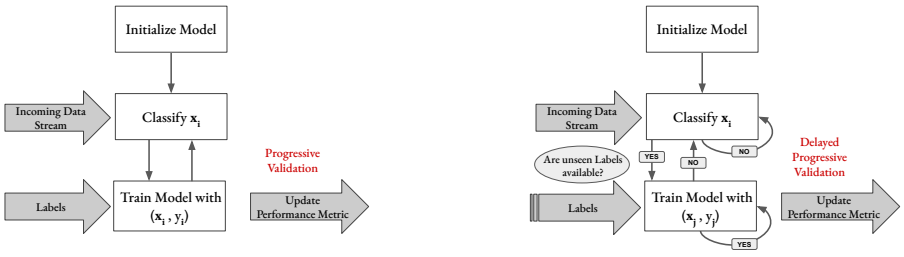
A and V stand for arousal and valence respectively, and $Y_{j,dimension}$ represents AS label by the participant after each video of two timestamps, t_{start} and t_{end}Ône label tuple $\mathcal{L}_{j,dimension}$ per video and dimension was sent from the PC running the PsychoPy experiment to the PC 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 done 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.

2.4. Evaluation

2.4.1. Online learning and Progressive validation

This paper aims at building a classification pipeline from evolving data streams. Several different *online learning*, or *stream learning*, algorithms have been proposed in the literature to deal with evolving data streams in supervised [53,54], unsupervised [55] and semi-supervised settings [56]. The static data from AMIGOS and from Experiment I was streamed using a library for online learning: *river* [57]. *Progressive validation*, also called

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2023 doi:10.20944/preprints202301.0156.v1



(a) **(b)**

Figure 7. 7a Progressive validation incorporated into the basic flow of the training process ('test-then-train') of an online classifier in an immediate label setting. $(\mathbf{x}_i, y_i)$ represents an input feature vector and its corresponding label. 7b Evaluation incorporated into the basic flow of the training process of an online classifier when labels arrive delayed.

test-then-train evaluation [58] is used for model evaluation in the supervised immediate label setting, when the labels for all samples are present at processing time [59]. Figure 7a shows the training process of an online classifier including progressive validation. Every time the model sees a new sample $\mathbf{x}_i$ 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, $\mathbf{x}_{i+1}$ before seeing its label, y_{i+1} and, again, do the training and performance metric updating step. This continues as long as data is streamed to the model. In this way, all samples can be used for training as well as for validation without corrupting the performance evaluation.

In the Experimental setup II, the labels are available after the prediction opposing the 'immediate labelling setting' [54] described as progressive evaluation. Therefore, a *delayed progressive validation* is performed with the delayed labels, which is mostly the case for real-life scenario. Figure 7b depicts the delayed progressive validation procedure, where the samples are classified by the model until an unseen labels are available. However, the 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 [60]. Once the model has been updated with all available labels, the classification of further samples continues with the now updated model. This can, of course, be implemented in parallel, as well. These steps continue as long as there is incoming data.

2.5. Machine Learning Classifiers

In this paper, three different algorithms: Adaptive Random Forest (ARF) [54], Streaming Random Patches (SRP) [61], and Logistic Regression (LR), have been evaluated and compared on the AMIGOS dataset and the data from Experiment I to find the best performing setup for the live emotion classification conducted in Experiment II. The ARF and and the SRP with a Hoeffding Adaptive Tree (HAT) [62] 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 [56,61]. The LR was included in the comparison as a sort of naïve baseline model by training on mini-batches (with partial fit) of 1 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 [57] are used with default parameters if not specified otherwise.

 doi:10.20944/preprints202301.0156.v1

2.5.1. 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 [19], the commonly reported metrics *Accuracy*, and *F1-Score* will be disclosed in this work. They are defined as follows [63]:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

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

$$\tag{4}$$

Where TP denotes the number of true positives classified by the model, and TN is the number of true negatives classified by the model. Accordingly, FP stands for the number of false positives classified by the model, and FN signifies the number of false negatives classified by the model. 'The higher, the better', can be said for both accuracy and F1-Score, i.e., a perfect model has an accuracy of 1 (100%) and an F1-Score of 1.

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 over 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.

3. Results

3.1. Immediate Label Setting

In this paper, first the real-time emotion classification pipeline was built by immediate label setting first and applied to data from AMIGOS dataset and Experiment I. The data were streamed to preprossses and to extract features from tumbling windows with a window length of 1 second. 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 for both valence and arousal models. For evaluation, as mentioned earlier, ARF, SRP, and LR classifiers were employed on 1 second window. The setting of 5 tress, and 4 tress for SRP worked best for AMIGOS dataset and for Experiment I, respectively. ARF included 5 tress for both datasets. A subject-dependent model was trained with 10-fold cross-validation, and the performance were evaluated with progressive validation.

Considering the data of established quality, we first validated the classification pipeline on the AMIGOS dataset. Table 3 presents the mean total calculated average of the F1-Score and accuracy over all the subjects achieved by each classifier with respect to affect dimensions. As depicted in "gray", both evaluation matrices reaches to 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. [21] and reported approximately 50% of F1-Score with no accuracy score reported. Siddharth et al. [64] reports more than 70% F1-Score and accuracy, and Topic et al. [65] achieves the current benchmark for this dataset by reporting 90% accuracy. However, all the related work mentioned were were obtained by using a hold-out or k-fold cross-validation with offline classifiers and the available labels at training time.

Figure 8 presents the overall model performances for individual subjects to showcase the subject-wise distribution of the evaluation matrix. The mean F1-Score for the positive and negative class of valence and arousal recognition, respectively were shown only for ARF and SRP, since the LR showed poor performance in comparison. The consistent higher F1-score mostly between 0.7 and 0.95 with two outliers validates the emotion classification

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2023 doi:10.20944/preprints202301.0156.v1

Table 3. Comparison of mean F1-Scores and accuracy of Valence and Arousal recognition on the AMIGOS dataset for short videos over all participants for different classifiers. Gray color 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. [21]	0.576	0.592	NR	NR
Siddharth et al. [64]	0.8	0.74	0.83	0.791
Topic et al. [65]	NR	NR	0.874	0.905

pipeline built in this paper. Two outliers are visible from subjects 11 and 30 might be due to a label imbalance (high/low) in the data or a bad data quality.

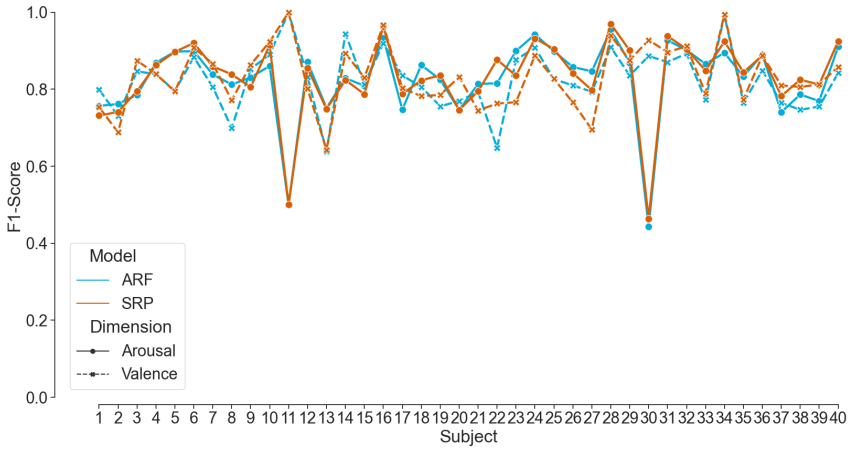


Figure 8. F1-Score for Valence and Arousal classification achieved by ARF and SRP per participant in the AMIGOS dataset.

After validating the classification pipeline on AMIGOS dataset, we evaluate the pipeline on the data from Experiment I. Table 4 presents the mean F1-Score of the subject-dependent models with the three classifiers for the positive and negative class of arousal and valence recognition, respectively. The F1-Score from both employed EEG device 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 [19, 21]. Furthermore, the two ensemble methods: ARF and SRP showed better performance with the mean F1-Score of more than 82% with no statistically significant difference ($p > 0.05$) in between. LR models showed poor performance similar to the performance on AMIGOS dataset. 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 than using Muse data. However, the differences are not statistically significant ($p > 0.05$) because 4 out of 11 cases for valence and 5 out of 11 cases for arousal recognition were showing the F1-Score for this dataset. Thus, the distribution of which device's data leads to the best performance per subject is actually rather balanced on this dataset.

3.2. Effects of Window Size

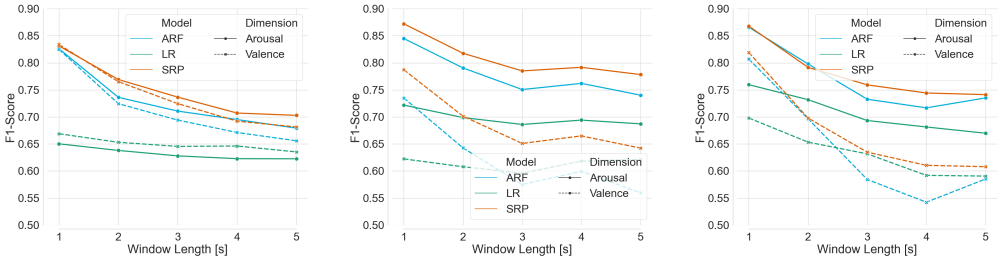
To find the optimum window length for online prediction, this paper extract features from different tumbling window length of $l \epsilon [1\,s, 2\,s, 3\,s, 4\,s, 5\,s]$ as depicted in Figure 6.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2023 doi:10.20944/preprints202301.0156.v1

Table 4. Comparison of mean F1-Scores of Arousal and Valence recognition per participant and device from Experiment 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	**0.902**	0.885	0.895	0.898	0.8	0.785
	4	0.836	0.794	0.838	**0.845**	0.793	0.604
	5	0.651	0.812	0.699	**0.827**	0.764	0.682
	6	0.836	0.843	0.863	**0.889**	0.771	0.62
	7	**0.958**	0.833	0.933	0.878	0.841	0.725
	8	0.889	0.749	**0.893**	0.783	0.683	0.584
	9	0.888	0.921	0.836	**0.931**	0.756	0.703
	10	**0.969**	0.903	0.951	0.915	0.816	0.898
	11	0.938	0.768	**0.955**	0.861	0.765	0.908
	12	0.864	0.871	**0.884**	0.878	0.669	0.697
	13	0.792	**0.913**	0.8	0.887	0.701	0.734
	Mean	0.866	0.845	0.868	**0.872**	0.76	0.722
Valence	3	0.837	**0.887**	0.811	0.876	0.716	0.712
	4	0.841	0.69	0.773	**0.859**	0.804	0.524
	5	0.546	0.734	0.639	0.748	**0.781**	0.58
	6	0.713	0.687	**0.785**	0.778	0.73	0.393
	7	**0.935**	0.666	0.926	0.757	0.776	0.616
	8	0.813	0.551	**0.819**	0.623	0.594	0.444
	9	0.812	0.844	0.721	**0.863**	0.72	0.561
	10	**0.982**	0.859	0.979	0.871	0.74	0.874
	11	0.924	0.653	**0.957**	0.811	0.64	0.884
	12	0.889	0.756	**0.914**	0.784	0.633	0.663
	13	0.584	**0.826**	0.6	0.775	0.543	0.595
	Mean	0.807	0.735	**0.819**	0.787	0.698	0.622

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2023 doi:10.20944/preprints202301.0156.v1



(a) AMIGOS. **(b)** Experiment I, Muse. **(c)** Experiment I, Crown.

Figure 9. Mean F1-Score achieved by ARF, SRP, and LR over the whole dataset for both affect dimension with respect to window length.

As detailed in Section 2.3, 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. With 10-fold cross-validation and progressive validation, the mean F1-Scores from ARF and SRP classifiers are depicted for AMIGOS dataset and datasets from Muse (Figure 9b) and Crown (in Figure 9c) from Experiment I. The Figure 6 shows that the best predictive performance was achieved with a window length of 1 second irrespective of the affect dimensions, classifiers and devices. Moreover, in most cases the classification performance is decreasing with increasing window sizes emphasizing the need of 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.

3.3. Delayed Label Setting: Live Classification

In order to validate the streaming setup of the emotion classification pipeline from Experiment I, live predictions and live online training was performed in Experiment II. The participants wore the same EEG device for both parts of the experiment: participant 14 and 17 wore the Muse headband and participant 15 and 16 wore the Crown. For each participant, an ARF with 4 trees was trained on the data recorded in part 1 of the experiment using a window length of 1 second and progressive delayed validation. With the pre-trained model, live predictions were performed with the data streaming in the part 2 of Experiment II. The prediction is 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 predictions in 73% for arousal and 60% for valence. However, most of the reported accuracy in Table 5 barely reached chance level. The lower predictive performance led us to investigate more on the delayed labels. To imitate production settings, we induced delay on the into the pipeline and applied progressive delayed validation on the subject-dependent model from Experiment I. Since the data from Experiment I was not a live stream, the model was updated with the true label for a sample after it had seen the next 86 samples i.e., the mean length of the video stimuli was 86 seconds. 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 4 trees and a window length of 1 s. The F1-score for individual participant reached to 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 verily compared to Table 4), when a delay is induced. However, the findings justifies the poor performance in the live settings and validates 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. [21] by 4.5% with a immediate label settings. However, the results

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2023 doi:10.20944/preprints202301.0156.v1

reported by Siddharth et al. [64], and Topic et al. [65] outperforms with the immediate labels.

Table 5. F1-Score and accuracy for the live affect classification in Experiment II (part 2). Subject 14 & 17 wore Muse, subject 15 & 16 wore the Crown for data collection.

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

Table 6. Mean F1-Scores for Valence and Arousal recognition of Experiment I, relayed per participant and device. Obtained using ARF (with 4 trees), a window length of 1 second, and progressive delayed validation with a label delay of 86 seconds. 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	**0.767**	0.561	**0.784**	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

4. 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 the existing data of ensured quality from the state-of-the-art AMIGOS [21] dataset. By streaming the recorded data to the pipeline, the mean F1-Score achieves 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 by [64] using the same dataset. Topic et al. [65] shows a better performance; however, due to the reported complex setup and computationally costly methods, the system is unsuitable for real-time emotion. 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 knowledge, no other work tested an online EEG-based emotion classification framework on the published AMIGOS dataset.

Secondly, a similar framework from the AMIGOS dataset 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

 doi:10.20944/preprints202301.0156.v1

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. 11 participants took part in Experiment I, where EEG data was recorded while watching 16 emotion elicitation videos. The pre-recorded data is streamed to the pipeline and showed a mean F1-Score of more than 82% with ARF and SRP classifiers using progressive validation. The finding validates the competence of the pipeline on the challenging dataset coming 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 verifies that the run-time on the training step of the pipeline of ARF is less than that of SRP, concluding to use of ARF 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 second is chosen for further analysis. Although the two employed consumer-grade devices possess a different number of sensors at contrasting positions, there were no statistically significant differences between the achieved performance scores on their respective data found. Therefore, we used both devices for live prediction, and the pipeline was applied to a live incoming data stream in Experiment II with the above-mentioned features of the model. In the first part of the experiment, the model is trained with the immediate labels from the EEG data stream. In the second part, the model is used to predict affect dimensions while the labels are available after a delay of the video length. The model is continuously updated whenever a new label is available. The performance scores achieved during the live classification with delayed labels are much lower than with immediate labels in Experiment I, motivating to induce artificial delay to the data stream from Experiment I. The results are compatible with the results from the live prediction. The literature reports better results for real-time emotion classification frameworks [23,24,26] 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 the realistic production scenario from daily life with the possibility of including further modifications in future work.

As a future work, the selected stimuli can be shortened to reduce the delay of the incoming labels so that the model is updated more frequently. Otherwise, multiple intermediate labels can also be included in the study design to ensure the inclusion of short time emotions felt while watching the movies. Furthermore, more dynamic preprocessing of the data can be included with feature selection algorithms for better 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 more high arousal ratings is also visible in the AMIGOS dataset, albeit not as intensely (62.5% high arousal ratings). In contrast, Betella et al. [46] found "a general desensitization towards highly arousing content" in participants. The underrepresented class can be upsampled in the model training in the future, or the basic emotions can be classified instead of arousal and valance, solving a multiclass problem [66]. Including more participants in the future for live prediction, the prediction can be visible to the participant as well to include neurofeedback. It will also be interesting to see if the predictive performance improves by utilizing additional modalities other than EEG, for example, Heart rate, Electrodermal activity [19,22,28].

Author Contributions: Conceptualization, Sidratul Moontaha, Franziska Schumann and Bert Arnrich; Data curation, Franziska Schumann; Formal analysis, Franziska Schumann; Resources, Sidratul Moontaha; Visualization, Sidratul Moontaha and Franziska Schumann; Writing—original draft preparation, Sidratul Moontaha; Writing—review and editing, Sidratul Moontaha, Franziska Schumann and Bert Arnrich; Supervision, Bert Arnrich. 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

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. The source code used for analysis in this study can be found athttps://github.com/HPI-CH/EEGEMO.

Acknowledgments: We appreciate the contribution of all the participants who participated for the sake of science. We also acknowledge the researchers from 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 And Negative Affect Schedules
PSD	Power Spectral Density
SRP	Streaming Random Patches

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2023 doi:10.20944/preprints202301.0156.v1

Appendix A

Input: Unlabelled EEG data stream $\mathcal{S} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i, ...\}$
Stream of true class labels including corresponding stimulus start- and end-times $\mathcal{L} = \{(\mathcal{Y}_j, \text{startTime}_j, \text{endTime}_j), (\mathcal{Y}_{j+1}, ...), ...\}$
Sampling frequency 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-timestaps(startTime, endTime,
             extractedData);
            matchedPredictions ← match-timestaps(startTime, endTime,
             predictions);
            for index in length(matchedWindows) do
                performance-metric-update(Y, matchedPredictions[index]);
                train-one(model, Y, matchedWindows[index]);
```

Algorithm 1: Live Emotion Classification from an EEG Stream

References

1. Picard, R.W. Affective Computing. *M.I.T Media Laboratory Perceptual Computing Section Technical Report* **1995**.
2. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE* **2001**, *18*, 32 – 80.
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 Behavior* **2012**, *23*, 415–421.
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 Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 2013, pp. 53–58.
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 research* **2006**, *70*, 83–88.
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.
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 & Behavior* **2016**, *55*, 157–164.

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.
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. Association for Computing Machinery, 2020, PervasiveHealth '20, p. 384–387.
10. Levenson, R.; Lwi, S.; Brown, C.; Ford, B.; Otero, M.; Verstaen, A., Emotion. In *Handbook of Psychophysiology, Fourth Edition*; 2016; pp. 444–464.
11. Liu, H.; Zhang, Y.; Li, Y.; Kong, X. Review on Emotion Recognition Based on Electroencephalography. *Frontiers in Computational Neuroscience* **2021**, *15*.
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. *Frontiers in Neuroscience* **2017**, *11*.
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), 2018, pp. 795–800.
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), 2018, pp. 1–6.
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), 2019, pp. 694–697.
16. Hasnul, M.; Ab.Aziz, N.; Alelyani, S.; Mohana, M.; Abd Aziz, A. Electrocardiogram-Based Emotion Recognition Systems and Their Applications in Healthcare—A Review. *Sensors* **2021**, *21*, 5015.
17. Huang, X.; Kortelainen, J.; Zhao, G.; Li, X.; Moilanen, A.; Seppänen, T.; Pietikäinen, M. Multi-modal Emotion Analysis from Facial Expressions and Electroencephalogram. *Computer Vision and Image Understanding* **2016**, *147*, 114–124.
18. Li, J.; Qiu, S.; Shen, Y.Y.; Liu, C.L.; He, H. Multisource Transfer Learning for Cross-Subject EEG Emotion Recognition. *IEEE Transactions on Cybernetics* **2020**, *50*, 3281–3293.
19. 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.
20. Horvat, M.; Dobrinić, M.; Novosel, M.; Jerčić, 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), 2018, pp. 1006–1010.
21. 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 Transactions on Affective Computing* **2021**, *12*, 479–493.
22. 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), 2020, pp. 576–579.
23. Bajada, J.; Bonello, F.B. Real-time EEG-based Emotion Recognition using Discrete Wavelet Transforms on Full and Reduced Channel Signals. *CoRR* **2021**, *abs/2110.05635*.
24. 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), 2014, pp. 3166–3171.
25. 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), 2020, pp. 2618–2625.
26. Lan, Z.; Sourina, O.; Wang, L.; Liu, Y. Real-time EEG-based emotion monitoring using stable features. *The Visual Computer* **2016**, *32*, 347–358.
27. Nandi, A.; Xhafa, F.; Subirats, L.; Fort, S. Real-Time Emotion Classification Using EEG Data Stream in E-Learning Contexts. *Sensors* **2021**, *21*, 1589.
28. 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 Transactions on Affective Computing* **2012**, *3*, 18–31.
29. Katsigiannis, S.; Ramzan, N. DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals from Wireless Low-cost Off-the-Shelf Devices. *IEEE Journal of Biomedical and Health Informatics* **2018**, *22*, 98–107.
30. Subramanian, R.; Wache, J.; Abadi, M.K.; Vieriu, R.L.; Winkler, S.; Sebe, N. ASCERTAIN: Emotion and Personality Recognition Using Commercial Sensors. *IEEE Transactions on Affective Computing* **2018**, *9*, 147–160.
31. Bradley, M.M.; Lang, P.J. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* **1994**, *25*, 49–59.
32. Watson, D.; Clark, L.A.; Tellegen, A. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* **1988**, *54 6*, 1063–70.
33. 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. *Electroencephalography and clinical neurophysiology* **1993**, *86 1*, 1–6.
34. Peirce, J.; Gray, J.; Simpson, S.; MacAskill, M.; Höchenberger, R.; Sogo, H.; Kastman, E.; Lindeløv, J. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods* **2019**, *51*.
35. Dan-Glauser, E.S.; Scherer, K.R. The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior Research Methods* **2011**, *43*, 468–477.

36. Kurdi, B.; Lozano, S.; Banaji, M. Introducing the Open Affective Standardized Image Set (OASIS). *Behavior Research Methods* **2016**, *49*.
37. Lang, P.J., B.M..C.B. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical report, 2008.
38. Panda, R.; Malheiro, R.; Paiva, R.P. Novel Audio Features for Music Emotion Recognition. *IEEE Transactions on Affective Computing* **2018**, *11*, 614 – 626.
39. Zhang, K.; Zhang, H.; Li, S.; Yang, C.; Sun, L. The PMEmo Dataset for Music Emotion Recognition. In Proceedings of the Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval; ACM: New York, NY, USA, 2018; ICMR '18, pp. 135–142.
40. 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 Transactions on Affective Computing* **2015**, *6*, 209–222.
41. Soleymani, M.; Lichtenauer, J.; Pun, T.; Pantic, M. A Multi-Modal Affective Database for Affect Recognition and Implicit Tagging **2012**. *3*, 42–55.
42. 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), 2019, pp. 3975–3979.
43. Mehrabian, A.; Russell, J.A. *An approach to environmental psychology*; The MIT Press, 1974.
44. Zhang, Y.; Fjeld, M. "I Am Told to Be Happy": An Exploration of Deep Learning in Affective Colormaps in Industrial Tomography. In Proceedings of the 2021 2nd International Conference on Artificial Intelligence and Information Systems. Association for Computing Machinery, 2021, ICAIIS 2021.
45. Breyer, B.; Bluemke, M. Deutsche Version der Positive and Negative Affect Schedule PANAS (GESIS Panel). *Zusammenstellung sozialwissenschaftlicher Items und Skalen* **2016**.
46. Betella, A.; Verschure, P. The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions. *PLoS ONE* **2016**, *11*.
47. Jiang, X.; Bian, G.B.; Tian, Z. Removal of Artifacts from EEG Signals: A Review. *Sensors* **2019**, *19*.
48. Sörnmo, L.; Laguna, P. Chapter 3 - EEG Signal Processing. In *Bioelectrical Signal Processing in Cardiac and Neurological Applications*; Sörnmo, L.; Laguna, P., Eds.; Biomedical Engineering, Academic Press: Burlington, 2005; pp. 55–179.
49. Akwei-Sekyere, S. Powerline noise elimination in neural signals via blind source separation and wavelet analysis. *PeerJ PrePrints* **2014**, *3*.
50. Sweeney, K.; Ward, T.; Mcloone, S. Artifact Removal in Physiological Signals-Practices and Possibilities. *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society* **2012**, *16*, 488–500.
51. Yao, D.; Qin, Y.; Hu, S.; Dong, l.; Vega, M.; Sosa, P. Which Reference Should We Use for EEG and ERP practice? *Brain topography* **2019**, *32*, 530–549.
52. 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 Transactions on audio and electroacoustics* **1967**, *15*, 70–73.
53. Bifet, A.; Holmes, G.; Pfahringer, B. Leveraging Bagging for Evolving Data Streams. In Proceedings of the ECML PKDD, 2010, pp. 135–150.
54. Gomes, H.M.; Bifet, A.; Read, J.; Barddal, J.P.; Enembreck, F.; Pfahringer, B.; Holmes, G.; Abdessalem, T. Adaptive random forests for evolving data stream classification. *Machine Learning* **2017**, *106*, 1469–1495.
55. Barddal, J.P.; Gomes, H.M.; Enembreck, F. SNCStream: a social network-based data stream clustering algorithm. *Proceedings of the 30th Annual ACM Symposium on Applied Computing* **2015**, pp. 935–940.
56. Parker, B.; Khan, L. Detecting and Tracking Concept Class Drift and Emergence in Non-Stationary Fast Data Streams. In Proceedings of the AAAI, 2015.
57. 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. *Journal of Machine Learning Research* **2021**, *22*.
58. Grzenda, M.; Gomes, H.M.; Bifet, A. Delayed labelling evaluation for data streams. *Data Mining and Knowledge Discovery* **2020**, *34*, 1237–1266.
59. Blum, A.; Kalai, A.T.; Langford, J. Beating the hold-out: bounds for K-fold and progressive cross-validation. In Proceedings of the COLT '99, 1999.
60. 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. 2013, p. 1222–1230.
61. 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), 2019, pp. 240–249.
62. Bifet, A.; Gavaldà, R. Adaptive Learning from Evolving Data Streams. In Proceedings of the Proceedings of the 8th International Symposium on Intelligent Data Analysis: Advances in Intelligent Data Analysis VIII; Springer-Verlag: Berlin, Heidelberg, 2009; IDA '09, p. 249–260.
63. Aggarwal, C.C. *Data Classification: Algorithms and Applications*, 1st ed.; Chapman & Hall/CRC, 2014; pp. 636–638.
64. Siddharth, S.; Jung, T.P.; Sejnowski, T.J. Utilizing Deep Learning Towards Multi-Modal Bio-Sensing and Vision-Based Affective Computing. *IEEE Transactions on Affective Computing* **2019**.

 doi:10.20944/preprints202301.0156.v1

65. Topic, A.; Russo, M. Emotion recognition based on EEG feature maps through deep learning network. *Engineering Science and Technology, an International Journal* **2021**, *24*.
66. Ekman, P.; Friesen, W. *Unmasking the face: A guide to recognizingemotions from facial clues*; Oxford: Prentice-Hall, 1975.