Towards Reducing Continuous Emotion Annotation Effort During Video Consumption: A Physiological Response Profiling Approach

SWARNALI BANIK, Department of Computer Science & Information Systems, BITS Pilani Goa, India
SOUGATA SEN, APPCAIR, Department of Computer Science & Information Systems, BITS Pilani Goa, India
SNEHANSHU SAHA, APPCAIR, Department of Computer Science & Information Systems, BITS Pilani Goa, and HappyMonk AI Labs, India
SURJYA GHOSH, APPCAIR, Department of Computer Science & Information Systems, BITS Pilani Goa, India

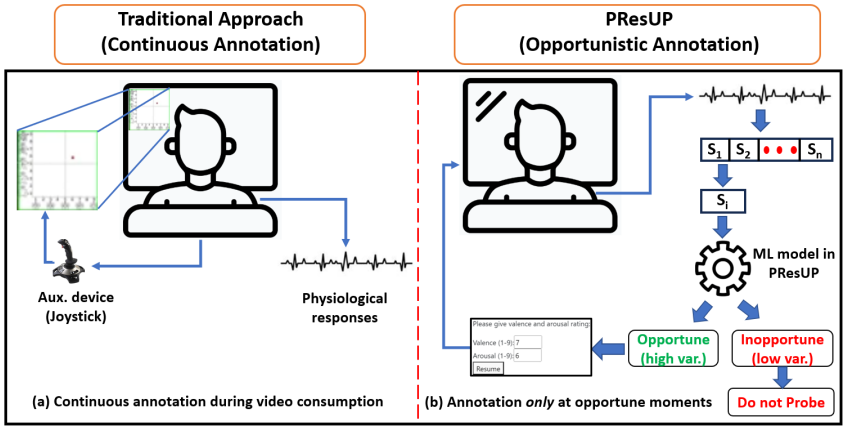


Fig. 1. Physiological response profile based opportunistic emotion annotation (PResUP) framework during video consumption scenario. (a) In traditional approach (absence of opportunistic annotation) during video watching, the user annotates *continuously* using an auxiliary device (e.g., joystick) and physiological signals also get recorded. (b) In the PResUP framework, the user is probed *opportunistically* for emotion annotation (self-report). The user annotates *only* those segments, which indicate major change in physiological response. So, the user does not need to annotate continuously, and therefore, survey fatigue is reduced without disrupting the viewing experience.

Emotion-aware video applications (e.g., gaming, online meetings, online tutoring) strive to moderate the content presentations for a more engaging and improved user experience. These services typically deploy a machine-learning model that continuously

Authors' Contact Information: Swarnali Banik, Department of Computer Science & Information Systems, BITS Pilani Goa, India, p20210016@goa.bits-pilani.ac.in; Sougata Sen, APPCAIR, Department of Computer Science & Information Systems, BITS Pilani Goa, India, sougatas@goa.bits-pilani.ac.in; Snehanshu Saha, APPCAIR, Department of Computer Science & Information Systems, BITS Pilani Goa, and HappyMonk AI Labs, India, snehanshus@goa.bits-pilani.ac.in; Surjya Ghosh, APPCAIR, Department of Computer Science & Information Systems, BITS Pilani Goa, India, surjyag@goa.bits-pilani.ac.in.

ACM 2474-9567/2024/9-ART91
https://doi.org/10.1145/3678569

infers the user's emotion (based on different physiological signals, facial expressions, etc.) to adapt the content delivery. Therefore, to train such models, the emotion ground truth labels also need to be collected continuously. Typically, those are collated as emotion self-reports from users in a continuous manner (using an auxiliary device such as a joystick) when they watch some videos. This process of continuous emotion annotation not only increases the cognitive load and survey fatigue but also significantly deteriorates the viewing experience. To address this problem, we propose a framework, PResUP that probes a user for emotion self-reports *opportunistically* based on the physiological response variations of the user. Specifically, the framework implements a sequence of phases - (a) user profile construction based on physiological responses, (b) similar user clustering based on the user profile, and (c) training a parameterized activation-guided LSTM (Long short-term memory) model by sharing data among similar users to detect the opportune self-report collection moments. All these steps together help to reduce the continuous emotion annotation overhead by probing at the opportune moments without compromising the annotation quality. We evaluated PResUP by conducting a user study (N=36) during which the participants watched eight videos, and their physiological responses and continuous emotion self-reports were recorded. The key results from this evaluation reveal that PResUP reduces the annotation overhead by reducing the probing rate by an average of 34.80% and detects the opportune probing moments with an average TPR of 80.07% without compromising the annotation quality. Motivated by these findings, we deployed PResUP by performing a follow-up user study (N=18). In this deployment scenario, we also obtained similar performance in terms of the probing rate reduction (average reduction of 38.05%), and opportune moment detection performance (average TPR of 82.26%). These findings underscore the utility of PResUP in reducing the continuous emotion annotation effort during video consumption.

CCS Concepts: • **Human-centered computing → Interactive systems and tools**; *Empirical studies in HCI*; Contextual design.

Additional Key Words and Phrases: Continuous emotion annotation, Physiological response, User Profile

ACM Reference Format:
Swarnali Banik, Sougata Sen, Snehanshu Saha, and Surjya Ghosh. 2024. Towards Reducing Continuous Emotion Annotation Effort During Video Consumption: A Physiological Response Profiling Approach. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 91 (September 2024), 32 pages. https://doi.org/10.1145/3678569

1 Introduction

In recent times, many video-based applications (e.g., online meeting platforms [43], online tutoring apps [31], entertainment platforms [24]) aim to moderate the content flow based on user emotion (e.g., stress, anxiety) to improve user engagement. For example, in case of online video lectures, if the stress level of the student is high, the lecture delivery can be moderated. These applications generally use machine learning models that infer emotions continuously based on the variations in the physiological responses as users watch the video. Therefore, to train such fine-grain emotion inference models, the emotion ground truth labels must also be collected continuously [70]. However, as emotion ground truth labels are typically collected as self-reports, the continuous emotion annotation process during video consumption becomes challenging because the users need to focus on two tasks simultaneously – (i) the users must watch the video, and (ii) they must provide the emotion self-reports. As a result, the viewing experience degrades, and the cognitive load increases. Therefore, efficient approaches for collecting continuous emotion annotations are essential.

In the existing literature, researchers primarily use an auxiliary device (e.g., mouse [14], joystick [55], or similar devices) to collect annotations continuously while the participants watch the videos. For example, in the CASE (*Continuously Annotated Signals of Emotion*) dataset [56], participants continuously provided emotion annotation (*valence* and *arousal* based on the Circumplex Model of emotion [50]) using a joystick. Works such as FEELTRACE [14] also collected continuous emotion annotations using similar devices. But these approaches have the same underlying challenge – concentrating on two tasks (video consumption and emotion annotation) at the same time. To address this challenge of simultaneously focusing on two tasks, Park et al. proposed a *post-interaction*, time-based approach of collecting annotations every 5-second interval for debate sessions (based

on audio-visual recordings) [47]. However, the challenge for this approach is the possibility of recall bias due to the post-interaction nature [35] and collecting significant annotations for a long session.

Several key intuitions may help to address the aforementioned challenges. First, human emotion persists for some time once experienced [63], commonly known as the persistence period of an emotion. Additionally, in a video, the rate of change of emotional content may not be very frequent (i.e., every frame may not depict a completely different emotion). Both these factors suggest that it may not be necessary to collect users' emotion ratings continuously. Second, the rich literature on emotion recognition suggests that the physiological signals change once the user experiences an emotion [19, 57]. Therefore, it may be possible to detect these moments of variations in the physiological responses and collect the emotion self-reports only at these *opportune* moments. Third, as the physiological signals are essentially time series data, it may be possible to detect these opportune moments using state-of-the-art sequential networks (e.g., LSTM - Long short-term memory) that are proficient in these tasks [34, 38].

We, in this paper, propose the PResUP framework for opportunistic continuous emotion annotation based on the physiological response profile of users leveraging the aforementioned intuitions. Based on these intuitions, we identify the specific video segments that cause variations in the physiological responses and collect emotion annotation only at these segments to avoid to continuous annotation. The framework operates in three phases (Section 5). First, it creates the user profile quantifying the variations in the physiological responses during different video segments. Intuitively, higher the variation in the physiological responses, more opportune the segment is for emotion self-report collection (as emotional stimuli influences the physiological responses [57]). The physiological response profile contains the summary statistics of user's physiological response behavior during the opportune and inopportune probing moments, thus provides an overview of user behavior at these moments. Second, we identify a group of similar users using k-means algorithm [30] based on the physiological response profile similarity. In the third (or final) phase, we enable data sharing among similar users to train a parameterized LSTM (p-LSTM) model (a variant of LSTM with parameterized Elliott activation function [59]) to detect the opportune moments for emotion self-report collection. The process of clustering similar users and sharing data among them helps to train the p-LSTM model with a larger pool of data for better performance. Fig. 1 presents the schematic diagrams to differentiate PResUP framework from the traditional continuous emotion annotation approaches. Unlike traditional approaches of continuously collecting emotion self-report during video consumption, we probe the user only during those moments when there is significant variation in physiological response (i.e., at the opportune moments); thus helping to reduce continuous annotation overhead.

We performed a lab-based user study (N=36, Section 3) to validate the hypothesis that emotion variations cause physiological response variations; therefore, the moments of physiological response variations can be considered for opportunistic annotation. During the study, the users watched a set of emotion stimuli videos (Table 1), provided emotion ratings (valence and arousal as per the Circumplex model of emotion [50]) continuously, and passively recorded their physiological responses (heart rate, galvanic skin response (GSR)) using the experiment apparatus (Section 3.1) developed for the study. We split the collected sensor data into small windows, computed the physiological response variations (in these segments using the RuLSIF algorithm [40]), and tagged a segment as opportune (or not) based on the degree of variation in physiological response. A detailed analysis of the dataset proves the existence of a causal relationship (Section 4) between emotion variation and physiological signal change, which reinforces that the opportune segments are ideal for probing the user for emotion self-reports. Motivated by this finding, we apply the PResUP framework on this dataset to identify the segments for opportunistic annotation. The empirical evaluation (Section 6) on this dataset reveals that PResUP reduces the average probing rate by 34.80% (with respect to baselines), detects the opportune probing moments with high accuracy (True Positive Rate: 80.07%, and False Positive Rate: 9.30%), and yet maintains the emotion annotation quality as observed in continuous annotations.

Although promising, the results of the previous study do not showcase PResUP's performance in a real-world deployment scenario. So, we performed a follow-up user study (N=18, Section 7) deploying the framework. In specific, we integrated the opportune moment detection model with the annotation application so that it detects in real-time whether a particular video segment is opportune (or not) for self-report collection and accordingly probes the user for annotation. As a result, this application alleviates the need for continuous annotation and collects annotations only at the opportune moments, thereby reducing the annotation effort. The empirical evaluation (Section 8) on this dataset (collected from this study) reveals that PResUP reduces the probing rate by 38.05% (on average) and detects the opportune moments accurately (average TPR: 82.26%, average FPR: 9.01%). A post-study qualitative survey also underscores the low interruption and high usability experience of the application. In summary, the major contributions of this paper are as follows:

- We propose a framework PResUP (Section 5) that reduces the continuous emotion annotation effort during video consumption. It encompasses a parameterized LSTM (p-LSTM) network, which leverages the physiological response pattern to identify the opportune moments for self-report collection, thus reducing the burden of continuous emotion annotation. We demonstrate that in spite of the reduction in the number of probes to be answered by the user, there is no significant difference in the quality of the emotion annotations.
- We propose a physiological response based user profile construction approach (Section 5.1). It allows to identify users with similar physiological response behavior (Section 5.2) so that the p-LSTM network embedded in the PResUP framework can be trained more efficiently (Section 5.3). The empirical evaluation demonstrates that training the network by sharing data among similar users helps to identify the opportune probing moments more accurately.
- We evaluate PResUP (Section 6) based on the dataset gathered from a real-world user study. The evaluation results on this dataset demonstrate a reduction in average probing rate (34.80%) and a high performance in detecting the opportune probing moments (avg. TPR: 80.07%, avg. FPR: 9.30%) without compromising the annotation quality. Moreover, we discussed the generalizability (Section 9.1.3) of the PResUP by demonstrating its superior performance over the baselines on couple of public datasets.
- Finally, we demonstrated the utility of PResUP by deploying it (Section 7, 8) and found that it detects the opportune moments accurately (avg. TPR: 82.26%, avg. FPR: 9.01%) and reduces the continuous emotion annotation effort substantially (avg. reduction: 38.05%).

These findings underscore the efficacy of the PResUP framework in reducing the continuous emotion annotation overhead without compromising the annotation quality during video consumption.

2 Related Works

In this section, first, we discuss the existing approaches for emotion annotation using self-reports and the continuous video annotation strategies. Later, we highlight the relationship between physiological responses and emotion from existing literature and discuss the related works on user profile creation based on physiological response behavior. Finally, we summarize how each of these aspects helps to develop the opportunistic annotation strategy to reduce the continuous annotation overhead.

2.1 Emotion Annotation using Self-report

In the existing literature, the emotion annotations are typically collected as self-reports using survey questionnaires due to the subjective nature of human emotion [37]. The most widely adopted approach for self-report collection is the post-interaction or post-stimuli one, where the participants provide emotion self-reports based on a standard scale (e.g., Self-assessment Manikin (SAM) [8]) after watching the video. As a result, these approaches fail to capture intra-video emotion variations due to the post-interaction annotation.

2.1.1 Continuous Emotion Annotation. To address this problem, researchers use continuous emotion annotation strategies, where participants provide emotion annotations using a mouse or joystick, as they watch the video. For example, earlier works such as Feeltrace [14], GTrace [15], DARMA [26] require users to continuously input the emotions as they watch the videos. Similarly, in the CASE (*Continuously Annotated Signals of Emotion*) dataset [56], participants continuously provided emotion annotation (*valence* and *arousal* based on the Circumplex Model of emotion [50]) using a joystick. Drawing on these set of works, Zhang et al. proposed a continuous emotion annotation wheel based on the Circumplex model of emotion so that the users can annotate the emotions on a mobile screen during video watching [70]. However, the challenges with these approaches are - (i) they require the users to use an auxiliary device for emotion annotation, (ii) due to the continuous nature of emotion annotation and video consumption (in parallel), mental workload increases and the viewing experience degrades.

2.1.2 Retrospective Emotion Annotation. To address the challenges of concentrating on two tasks at the same time, researchers proposed to annotate the segments retrospectively. For example, Park et al. [47] collected emotion annotations (valence, arousal on a scale of 1 to 5) at every 5-second interval based on the debate performed by two subjects. To reduce the burden further, Bota et al. [7] suggested to identify *only* those segments, which produced major variations (above a threshold) in the EDA signals and collect annotation from the users for these segments. However, these approaches require users to watch videos multiple times, which becomes more time-consuming for longer videos, and the need for users to recall previous emotions associated with the specific video segments during initial viewing.

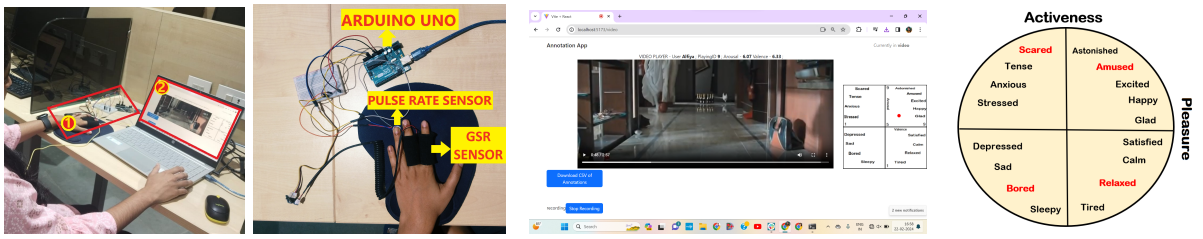
2.2 Emotion and Physiological Responses

The rich literature in affective computing highlights a strong relationship between human emotion and different physiological signals [57]. For example, physiological signals like EEG, ECG have been explored extensively to detect different types of emotions or different levels of valence (or arousal) [2, 33, 39, 44]. Similarly, other physiological signals such as Galvanic skin response (GSR) or Electrodermal activity (EDA), Electromyography (EMG), Respiratory signal (RSP) have also found to be effective in detecting different emotions (or different levels of valence/arousal) [58, 68, 69]. Moreover, many earlier works showed the combination of different types of physiological signals are also effective in detecting different emotions [18, 25, 60].

Typically, the emotion inference pipeline from the physiological responses uses a machine learning model, which either involves manual feature extraction [3, 57] or neural network [51, 61]. Example of manually extracted features include Fast Fourier Transform (FFT) on EEG signal, Wavelet transform (WT) on EEG, ECG, RSP signals [27, 45]. Similarly, recent advances in the neural network automatically extract representations (from raw physiological signals) that correlate well with the emotions [52, 65]. However, both the approaches highlight that as human emotions vary, there are noticeable changes in the physiological responses [57]. We leverage such variations in the physiological signals to detect the opportune probing moments for emotion self-report collection rather than asking users to continuously input their emotions.

2.3 Physiological Response based User Profile

The existing literature on social psychology indicates that physiological responses vary based on the user characteristics (e.g., Personality traits) [64]. For example, extraverts require more external stimulation than introverts, and that neurotics are aroused more easily [21]. Therefore, it may be possible to group users based on the physiological response profile, defined as the physiological response behavior of a user based on different types of stimuli [46]. Although different types of user profiles (e.g., based on user preference [16, 23]) are explored widely in the existing literature for various tasks (e.g., recommender systems [6]), the application of physiological response profile is relatively underexplored.



(a) User study setup **(b)** Sensor setup **(c)** The "Annotate App" **(d)** Circumplex model [50]

Fig. 2. The overview of the user study - (a) study setup (including the sensors and the UI for video consumption and annotation using keyboard arrows) (b) Sensor configuration for physiological response (GSR, HR) collection. (c) "Annotate App" UI for video consumption and annotation (d) the Circumplex model of emotion [50], which guides the emotion annotation in a 2D plane (valence, arousal).

The recent developments highlight that neural networks usually require a large amount of training data for superior performance [32, 36]. One possible approach to counter the data scarcity is to share data among *similar* users. For example, in recommender system design, often the data among *similar* users are shared for better performance [28]. Recently, Hari et al. demonstrated that affective profile similarity-based data sharing can lead to better emotion recognition [29]. Motivated by these examples, we also considered to construct physiological response profile to enable data sharing among similar users so that the neural network models can detect the opportune probing moments accurately.

Key Takeaways: In summary, probing a user opportunistically for emotion self-report may be an alternative for continuous emotion annotation during video consumption. While machine learning models can be used to detect the opportune moments, they may require large training data, which can be overcome by sharing data among users with similar physiological response profiles. However, such an approach has not been investigated in the case of continuous emotion annotation, which we undertake in this paper.

3 User Study I: Physiological Response and Continuous Annotation Collection

The objective of this user study is to collect physiological responses and continuous emotion annotations from the participants during their video consumption. The collected dataset from this study is used for the following purposes - (a) to validate that emotion variations cause the physiological response variation (Section 4) (b) To make design choices while developing the PResUP framework (Section 5).

In this section, we discuss the user study including experiment apparatus, study procedure, pre-processing and the collected dataset. This work has been approved by our institute's Human Ethics Committee (HEC), and we have obtained the IRB approval (HEC/BPGC/2023/005) prior to the user study.

3.1 Experiment Apparatus

The experiment apparatus consists of two key components (Fig. 2a). The *first* component includes a pulse rate sensor (HW-827, World Famous Electronics LLC) and a galvanic skin response (GSR V1.2, Seeed Studio Grove) sensor connected to the GPIO pins of an Arduino Uno board. These sensors record the heart rate (HR) and skin conductance respectively at a sampling frequency of 10 Hz (for each sensor). The sensor data is transferred to a connected laptop via the serial port. The sensor configuration is shown in Fig. 2b. These sensors were selected strategically to ensure real-time measurements with precise and accurate data[1] while also maintaining affordability and can be implemented practically in a variety of situations [42].

[1]We compared the signal quality of these sensors with that of an Empatica Embrace Plus in Appendix A.1.

The *second* component is an UI ("Annotate App" in Fig. 2c) that displays the videos and allows users to provide emotion annotations continuously using the keyboard. The emotion annotations are collected based on the Circumplex model [50]. This model represents human emotion in two dimensional plane (valence, signifying pleasure and arousal, signifying activeness) divided into four quadrants (Fig. 2d). When the application is launched, the cursor (represented by a red dot in circumplex model) starts at the origin point (Fig. 2c). As the video proceeds, the user moves the arrow keys to indicate experienced emotions (valence and arousal) on the keyboard. At a specific timestamp, the cursor position assigns valence and arousal ratings to video content. Launching the annotate app triggers the Arduino to passively record physiological signals (HR and GSR) while users annotate their emotions during video playback.

3.2 Study Procedure

We recruited 36 participants (18F, 18M) aged between 20 and 40 years from our university. During the study, each participant watched 8 videos and recorded their emotion continuously via the Annotate App. The physiological signals were recorded passively using the sensor setup. We selected these videos as the embedded emotion of these videos (see Table 1) cover all quadrants (one emotion from every quadrant) and have been used for similar tasks in earlier studies [55]. We present the video details with related emotions (amused, relaxed, bored, scared) of every video in Table 1.

Video id	Emotion	Valence	Arousal	Duration (in sec.)
1	amusing	med/high	med/high	185
2	amusing	med/high	med/high	173
3	boring	low	low	119
4	boring	low	low	160
5	relaxing	med/high	low	145
6	relaxing	med/high	low	147
7	scary	low	high	197
8	scary	low	high	144

Table 1. Details of the stimuli videos used in the user study. The same videos were used in earlier studies [55] for physiological response and continuous emotion annotations.

Every participant watched the videos in a predetermined random order and recorded the valence and arousal scores continuously on a scale of 1 to 9 using the keyboard arrow keys (the scores were automatically assigned based on the cursor position). During the annotations the users were not required the input the annotation values using the keyboard, rather they moved the arrow keys and accordingly the scores get assigned based on the cursor position. The annotation interface presents the circumplex plane containing the discrete emotion labels (e.g., amused, stressed, bored, relaxed), which helped the users to move the cursor accordingly to the direction of the perceived emotion. Based on the cursor position on the circumplex plane, we recorded the value of valence and arousal on a 9-point scale. As a result, although the annotations are collected on a 9-point scale, the users were reporting the emotions they felt. This annotation process allowed to reduce the cognitive load and guided the users to annotate correctly. A two-minute-long blue screen interleaved the videos to avoid the carry-over effect. The participants were explained about valence and arousal component of emotion. They were also instructed how to record emotions using the arrow keys during video consumption.

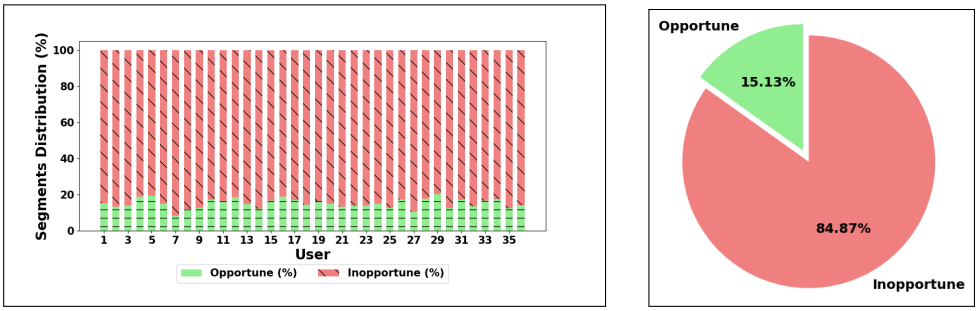
3.3 Data Pre-processing

We performed two data pre-processing steps on the raw data collected from the user study,

3.3.1 Segmenting Physiological Response. We segmented the physiological signals into 5-second fixed-size windows (based on the earlier works [1, 47]) for every participant and every video combination.

3.3.2 Labeling Opportune Moments. We labeled the segments as opportune (or not) based on the degree of variation of the physiological signals in a window - a higher variation in a window is the manifestation of the emotional change [1] during that time (hence opportune for probing user for emotion self-report).

We performed the following steps for each video. First, we compute the change point score between every two consecutive windows (5-second segment) by applying the RuLSIF algorithm [40]. A change point score denotes the degree of change in the time-series values (i.e., physiological responses) of two consecutive windows [5]. A high value of this indicates a major change in the physiological signals and therefore, indicator of emotional change. Second, we separate out the higher end ($> \mu + 3 * \sigma$) outliers, and perform k-means clustering (k=2) to cluster the remaining change point scores. We pick the cluster with higher centroid value and identify only those points within the cluster having value higher than the centroid. These points and the outlier points are marked together as opportune moments (as they indicate significant change in the physiological responses). A similar strategy has been adopted by Aditya et al. to identify significant changes in physiological signals [1]. We present these steps for labeling opportune moments as an algorithm (Algorithm 1) in Appendix A.2.



(a) User-wise segment distribution **(b)** Segment distribution (total)

Fig. 3. Distribution of opportune and inopportune segments (a) user-wise distribution, (b) distribution in the total dataset.

Total number of segments	8608
Total number of opportune segments	1303
Total number of inopportune segments	7305
Average (SD) number of segments per users	239 ± 1.29
Duration of every segment	5 Second
Duration of total sensor dataset	11.96 Hr.

Table 2. Final dataset details

3.4 Final Dataset Details

We obtained 8608 segments (5-second window) after the segmentation process from all the users. This dataset is equivalent to ≈ 12 hours of sensor and video data. Each user contributes to an average of 239 windows (SD:1.29). We did not observe a large variation in the number of segments for every user as they watched the same set of videos. We present the user-wise distribution of opportune and inopportune segments in Fig. 3a. In this case also, we observe the user-wise distribution of opportune and inopportune segments are similar. In total, we marked

15.13% segments as opportune and 84.87% segments as inopportune (Fig. 3b). We summarise the final dataset in Table 2.

4 Feasibility Analysis: Causality between Emotion and Physiological Response

In this section, we describe the feasibility study performed to test the hypothesis that emotion variations cause physiological signal variations. If this hypothesis is found to be true, then emotion variations can be captured by probing *only* at the opportune moments (denoted by substantial change in the physiological signals). We validate this hypothesis using the dataset collected from the user study I (Section 3.4).

4.1 Causality between Emotion and Physiological Response

Emotion variations are reflected in valence (or arousal) scores, whereas the physiological signal variations are manifested in the change point scores. Therefore, we investigate the causal relationship between valence-arousal scores and change point scores. In specific, we performed the Granger causality test [22] to verify if variation in valence and arousal causes variation in change point score. This test is a statistical hypothesis test for determining whether one time series (in this case emotion scores) is useful in forecasting another (in this case change point score). Mathematically it is expressed as follows. One time series X_t does "Granger-causes" another time series Y_t, if the past values of X_t help to predict the future values of Y_{t+1}. Granger causality or G-causality works on the principle of modeling the two processes X and Y as auto-regressive processes. Specifically, in order to determine if '*Y G-causes X*', the two models considered are: –

$$X(t) = \sum_{\tau=1}^{\infty}(p_\tau X(t-\tau)) + \sum_{\tau=1}^{\infty}(r_\tau Y(t-\tau)) + \varepsilon_c, \tag{1}$$

$$X(t) = \sum_{\tau=1}^{\infty}(q_\tau X(t-\tau)) + \varepsilon, \tag{2}$$

where t stands for time, p_τ, q_τ, r_τ are coefficients at a time lag of τ and $\varepsilon_c, \varepsilon$ are error terms. Covariance stationarity is assumed for both X and Y. Whether Y G-causes X (or not) can be predicted by the measure known as F-statistic which is the log ratio of the prediction error variances:

$$F_{Y\to X} = \ln \frac{var(\varepsilon)}{var(\varepsilon_c)}. \tag{3}$$

If the model represented by equation (1) is a better model for $X(t)$ than equation (2), then $\text{var}(\varepsilon_c) < \text{var}(\varepsilon)$ and $F_{Y\to X} > 0$, suggesting that *Y Granger causes X*. Even though G-causality uses the notion of autoregressive models for the variables, the generic nature of this modeling with minimal assumptions about the underlying mechanisms makes it a very popular choice in a wide range of disciplines [59, 66].

4.2 Causality Test Outcome

We perform this test twice (once for arousal and once for valence). We accumulate the arousal scores and the change point scores for every segment in two different lists and run the causality test (using the package[2]). The test outputs the following results ($F = 3.9048$, $\chi^2 = 3.9062$, p-val = 0.0481, df = 1). The null hypothesis is that arousal scores do not Granger cause the change point scores. However, as per the results ($p < 0.05$), we reject this null hypothesis, thus implying a causal relationship between arousal and change point scores. We repeat the same steps for valence. Like arousal, the null hypothesis is that valence scores do not Granger cause the change point scores. The test returns the following results ($F = 14.1973$, $\chi^2 = 14.2022$, p-val = 0.0002, df = 1), and therefore, we reject the null hypothesis (as $p < 0.05$). This outcome suggests that a causal relationship does exist between

[2] https://tinyurl.com/funzzup7

valence and change point scores. These findings highlight that emotional variations reflect in physiological signal changes (as emotion comprises both valence and arousal). Therefore, by probing at the opportune segments (i.e., points with high physiological response variation), we can capture the emotional changes. So, we aim to detect the opportune moments (using the PResUP framework) as discussed next.

5 PResUP Framework

In this section, we describe the PResUP framework. We present the overview of the PResUP framework in Fig. 4. The framework operates in three phases – (a) physiological response profile creation, (b) similar user identification, (c) opportune moment detection. In the profile construction phase, we create user profile based on the physiological responses observed in opportune and inopportune segments. Based on the physiological response profile similarity, we identify similar users using the k-means clustering algorithm. Once the similar users are identified, we share data among these users to train a parameterized LSTM model (for every cluster) for opportune moment detection. We discuss each of the phases in detail now.

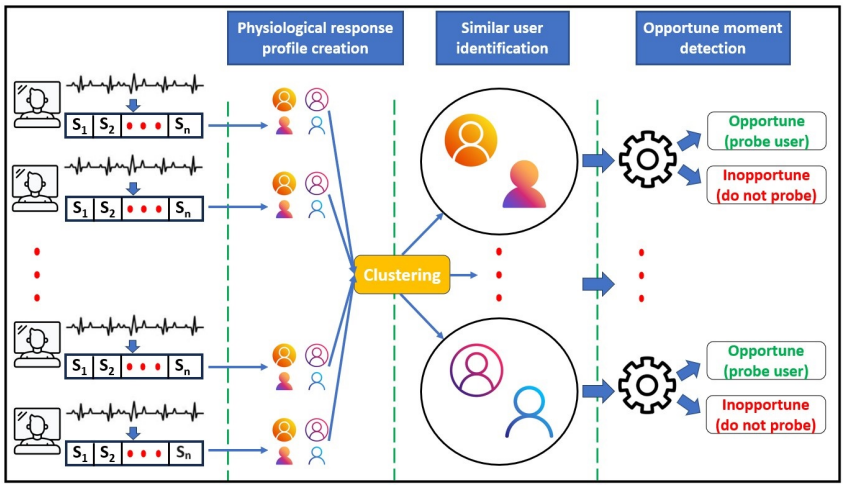


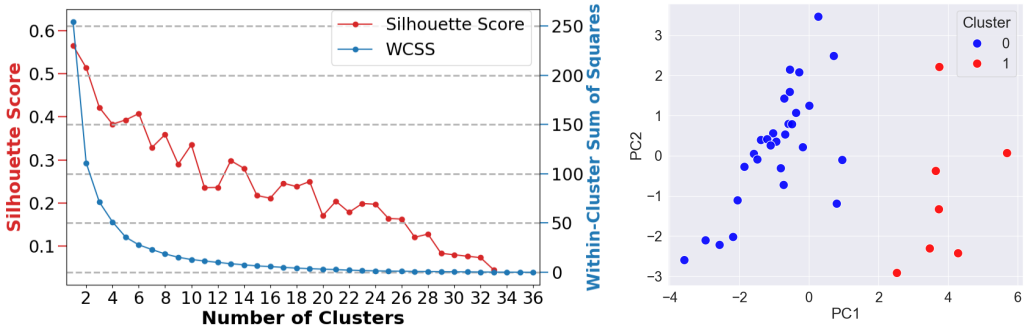
Fig. 4. Overview of the PResUP framework. It operates in three phases - (a) physiological response profile creation (b) similar user identification (c) opportune moment detection. In the first phase, during video consumption, the physiological responses are recorded that are used to create the response profile. In the second phase, the physiological response profile of the users are used to cluster the similar users. In the last phase, cluster-specific machine learning model (using p-LSTM) for opportune probing moment detection is constructed by sharing data among those users, who belong to the same cluster.

5.1 Physiological Response Profile Creation

The physiological response profile creation for a user is a two-step process. In the first step, we compute the change point scores between every two consecutive segments (s_i, s_{i+1}, where $1 \leq i \leq n-1$; n is the total number of segments). In the second step, we compute the mean (μ) and standard deviation (σ) of the change point scores for both opportune and inopportune segments. The user profile is therefore constructed as a four tuple $<\mu_{opp}, \sigma_{opp}, \mu_{inopp}, \sigma_{inopp}>$, where μ_{opp}, σ_{opp} indicate the mean and std. dev at the opportune segments, and μ_{inopp}, σ_{inopp} indicate the mean and std. dev at the inopportune segments. These steps are repeated for every user. Intuitively, change point scores indicate the variation in the physiological responses after applying the emotional stimuli. Therefore, profile creation by combining two types of change point scores can help to identify the group of users whose physiological response variations are similar during opportune and inopportune moments.

5.2 Similar User Identification

In this phase, we aim to identify the similar users based on the physiological response profile. We run the k-means algorithm on the response profile vectors for different values of k. We vary the value of k from 2 to 36 (as there were 36 users, Section 3.2), and note the cluster quality in Fig. 5a using Silhoutte score [49] and Within-Cluster Sum of Squares (WCSS) of elbow method [53]. We observe the highest value of the Silhouette score and the elbow pattern (in the plot) for k = 2. Therefore, we group the users into two clusters. To visualize the user groups, we perform PCA (Principal Component Analysis) [67] on the profile vectors and display the clusters in a 2D plane (Fig. 5b). Once we group the users into clusters, we proceed to construct the opportune moment detection model by sharing data among the similar users (i.e., users present in the same cluster).



(a) Silhoutte score and WCSS variation

(b) User clustering (k=2)

Fig. 5. Clustering users based on the physiological response profile vectors - (a) variation in Silhouette score and Within-Cluster Sum of Squares (WCSS) score for different cluster number (k) which reveals that number of optimal clusters is two (b) visualization of group of similar users for optimal number of cluster (k=2) after applying PCA on the profile vectors

Cluster #	Cluster Centroid ($\mu_{opp}, \sigma_{opp}, \mu_{inopp}, \sigma_{inopp}$)	Age Groups (age_group: #M, #F)	#of Males	#of Females
1	[7.690, 2.805, 6.053, 1.666]	(20-24: 7M, 5F), (25-29: 4M, 7F), (30-34: 2F), (35-40: 2M, 2F)	13	16
2	[5.709, 2.155, 1.661, 1.457]	(20-24: 1M, 1F), (25-29: 4M, 1F)	5	2

Table 3. Cluster analysis reveal that (i) the users with substantial different profile vectors are grouped in different clusters (observed in the different average (μ) values of the centroid) (ii) users in a cluster have similar profile vector (observed in the low SD (σ) values of the centroid) (iii) demographic details do not play a major role in differentiating the users (as users with different gender and age-group) are present in both clusters.

5.2.1 Cluster Analysis based on Physiological Response Profile. We perform cluster analysis based on the physiological response profile vectors and present the findings in Table 3. The key differentiating factor between the clusters is the physiological response profile behavior (captured using the response profile vectors). We observed that the centroids of the two clusters are far from each other (the values of the mean change point score for opportune and inopportune moments are very far $C_1 : \{\mu_{\text{opp}} = 7.690, \mu_{\text{inopp}} = 6.053\}; C_2 : \{\mu_{\text{opp}} = 5.709, \mu_{\text{inopp}} = 1.661\}$). We also noted that the physiological response properties are very similar as noted in the low std. deviation values ($C_1 : \{\sigma_{\text{opp}} = 2.805, \sigma_{\text{inopp}} = 1.666\}; C_2 : \{\sigma_{\text{opp}} = 2.155, \sigma_{\text{inopp}} = 1.457\}$) of the centroids. The demographic properties (age group, gender) do not play a differentiating role while grouping the users as each cluster has representation from both genders and varied age groups. Therefore, as the clustering is performed based on the physiological response properties of the users, it generalizes across demographic properties.

5.3 Opportune Moment Detection

To detect the opportune probing moments, we develop parameterized-LSTM (p-LSTM) [59] models by sharing data among users present in a cluster. The p-LSTM is similar to an LSTM network, where the activation functions of the LSTM gates (input, forget, and output) are replaced with parameterized Elliott activation function [20]. This activation function allows to capture dataset specific complex relationship and avoids the need of separate hyperparameter-tuning effort where the p values are learnt from data via gradient descent [59].

We show the model architecture in Fig. 6. Given a sequence of segments $\mathcal{X} = \langle \mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_T \rangle$ containing physiological responses, the model takes an input $\mathbf{x}_t = [cp_t, gsr_t, hr_t, v_{video}, a_{video}]^T$ at each step t, where cp_t is the change point score at time t, gsr_t is the difference in the mean value of GSR between segment t and $t-1$, hr_t is the difference in the mean value of HR between segment t and $t-1$; v_{video} and a_{video} denote the valence and arousal of the video respectively. These two values remain same in every segment for a given stimuli video depending on the valence and arousal of the video. For example, for stimuli video 1, the valence and arousal values are '1' as this video embeds the emotion amusing, which belongs to the first quadrant of the Circumplex model (Fig. 2d). We present the details of the stimuli videos in Table 1. Notably, the effectiveness of GSR (or EDA) and HR sensors have been demonstrated in many earlier works (e.g., [11], [17]) to infer emotional engagement. So we use GSR and HR values as input to the model.

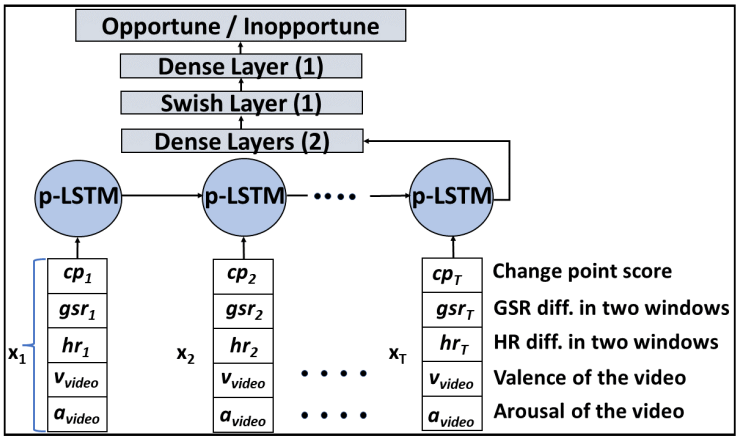


Fig. 6. p-LSTM-based architecture used in the PResUP framework for opportune probing moment detection

We incorporated parameterized Elliott activation functions (PEF) [20, 59] to replace the activation functions of the LSTM gates (input, forget, and output) in the p-LSTM architecture. The PEF is represented by

$$f(x) = \frac{px}{1 + |x|} \tag{4}$$

The first-order derivative of the function PEF is expressed as: $f'(x) = \frac{p}{(|x|+1)^2}$. This derivative becomes zero when the function equals the parameter p at the origin. Upon incorporating the PEF (Parameterized Elliott Function), the hidden state equation is given by: $h_t = O_t p_c PEF(C_t)$. By the chain rule,

$$\frac{\partial J}{\partial p_c} = \frac{\partial J}{\partial p_c} = \frac{\partial J}{\partial h_t} O_t * Elliott(C_t) \tag{5}$$

After each iteration, the p_c is updated by gradient descent ($p_c^{(n+1)} = p_c^n + \delta * \frac{\partial J}{\partial p_c}$). These modified activation functions introduce model parameters and enhance flexibility to capture complex data patterns (via adjusting slope, p) and relationships.

The usage of the parameterized Elliott activation function offers the following advantages - (a) as p in Eq. 4 is learnt from the dataset through back-propagation; if the dataset changes, the final activation form adjusts accordingly; thus removing the necessity for separate "parameter-tuning" efforts (b) PEF has a lower saturation rate, compared to the standard non-parameterized Elliott function (e.g., tanh, sigmoid) [59].

The proposed p-LSTM network consists of T cells[3]. The p-LSTM embedding is input to two fully connected dense layers with parameterized Elliott activation functions followed by a swish activation layer [48]. The swish activation function is similar to the ReLU (Rectified Linear Unit) activation function, but it introduces a non-linearity based on the sigmoid function. Swish improves p-LSTM over ReLU due to its non-monotonic nature [4]. The swish layer output is finally connected to a sigmoid function as it is used for the binary classification (i.e., opportune or not). Therefore, we use binary cross-entropy loss for training this architecture. We specify the hyperparameter details in Section 6.1.3. While training the model for a particular user u, we use data from *only* those users who belong to the same cluster as u (Fig. 5b, Section 5.2). We do not use any data from user u for training the model, rather it is used for evaluating the model. The same process is repeated for all the users.

6 Evaluation: User Study I

In this section, we evaluate the performance of PResUP in reducing the annotation effort (measured in terms of probing rate), detecting the opportune moments, and maintaining the annotation quality. Accordingly, we compare the performance of the PResUP with a set of baselines using the metrics defined below.

6.1 Experiment Setup

The experiment setup for evaluation of PResUP is as follows. First, we construct the physiological response profile of the users and cluster them. Next, for every user (u), we train the proposed p-LSTM model using data of other users present in that cluster only and test using the left-out user's (u) data. The process is repeated for each user.

6.1.1 Performance Metrics. We use the following metrics to evaluate the performance of the PResUP framework.

- **Probing Rate:** The number of video segments detected as opportune is the number of probes answered by a user (because the user will not be probed at the inopportune segments). We compute the average number of probes issued per video for every user as the probing rate.
- **Opportune Moment Detection Metrics:** We use the standard metrics to measure the probing moment detection performance. **True Positives (TP):** Opportune moments that are correctly identified. **False Positives (FP):** Inopportune moments that are identified as opportune. **True Negatives (TN):** Inopportune moments that are identified as Inopportune. **False Negatives (FN):** Opportune moments that are identified inopportune. We calculate: **True Positive Rate (Recall or Sensitivity) TPR(%)** = $\frac{TP}{TP+FN}$ x 100. **True Negative Rate (Specificity) TNR(%)** = $\frac{TN}{TN+FP}$ × 100. **False Positive Rate (Fall-out) FPR(%)** = $100 - TNR(\%)$. **Likelihood Ratio for opportune moments (LR+)** = $\frac{TPR(Sensitivity)}{FPR(1-Specificity)}$.
- **Annotation Quality:** We investigate if there is any statistically significant difference (significance level, $\alpha = 0.05$) in the mean valence (and mean arousal) between the ground truth annotations (continuous) and the annotations collected by probing at the opportune moments.

6.1.2 Baseline. We discuss the baselines below used for comparing the performance of the PResUP framework.

[3]We found (empirically) superior performance in detecting opportune moments for T=10. Therefore, we decide to use T=10.

- **Time-based Probing Strategy (TBS) [47]:** This baseline is inspired by earlier work on continuous emotion annotation [47]. In this approach, we probe the user at every 5 seconds. We select 5 sec. interval based on earlier work [10, 47]. Moreover, in our case, the duration of every segment is also 5 seconds. We compare PResUP with this baseline as it demonstrates the efficiency of the opportunistic probing over continuous (fixed interval) probing.
- **Random Probing Strategy (RPS):** In this approach, we randomly probed in some of the windows for annotation collection. We ran multiple iterations (n=20) of the baselines, and computed the average of different performance metrics.
- **Retrospective Probing Strategy (RePS):** In this baseline, we adopt the annotation method proposed by Bota et al. [7], focusing exclusively on the EDA signal. Similarly, we only consider the GSR signal while implementing this baseline. This work suggests to annotate only those segments in which the EDA value is above a threshold (0.01% of maximum EDA) of the minimum EDA. Accordingly, we annotate only those segments as opportune, which follows this rule. We compare the PResUP with this baseline to highlight the advantage of profile-based clustering over a rule-based annotation approach.
- **Feature-based Probing Strategy (FBS):** In this baseline, we train a classical machine learning model (not LSTM) by adopting the same approach (user profile creation, followed by clustering) as used in PResUP. It does not consider the sequence of segments, rather extracts the same features from the current segment for opportune moment detection. Notably, we implemented multiple algorithms (Random Forest [9], Support Vector Machine [13], XG Boost [12]). The results presented here are based on Random Forest as it returned the best performance. The comparison with this baseline underscores the utility of segment sequence in p-LSTM.
- **Personalized Probing Strategy (PPS):** This baseline implements a p-LSTM model like PResUP. However, the model is *personalized*. The model is trained using 80-20 split, where *initial* 80% physiological data of a user is used for training and the remaining 20% is used for testing. Notably, comparison with this baseline demonstrates the generalizability of PResUP, i.e, how well it can perform in absence of personal data during training.
- **Generalized Probing Strategy (GPS):** In this case, we train a p-LSTM network as used in PResUP. However, we do not create the user profile and identify the similar users (like PResUP). Rather, we adopt a Leave-one-subject-out cross-validation approach to evaluate this baseline. We compare PResUP with this baseline to showcase profile-based clustering returns superior performance instead of indiscriminately aggregating data from all users.
- **Age-based Probing Strategy (APS):** This baseline adopts same architecture like PResUP. However, in this case, clustering is performed based on the subject's age group. As there are four age groups (20-24, 25-29, 30-34, and 35-39 years) in the collected dataset, we created four clusters. To evaluate the model, we consider one user (test user) at-a-time and train the model using data from *only* those users present in the same cluster of the test user. The process is repeated for each user. Comparison of PResUP with this baseline demonstrates the superiority of physiological profile-based clustering over demographic-based (age) clustering.
- **Gender-based Probing Strategy (GBPS):** This is similar to the **APS** baseline with only difference is that the clustering is done based on the gender (male or female) of the participants. This helps to determine the efficiency of the physiological profile-based clustering over demographic-based (gender) clustering.
- **RNN-based Probing Strategy (RNNPS):** This baseline implements a recurrent neural network (RNN) layer employing the same method as PResUP (i.e., user profiling and user clustering). Two dense layers with ReLU activation are added, with the final layer facilitating binary classification using sigmoid activation. We present the hyperparameter details in Section 6.1.3.
- **GRU-based Probing Strategy (GRUPS):** This baseline employed a gated recurrent unit (GRU) architecture, followed by three dense layers with ReLU activation. Finally, ending with a single neuron output layer facilitated

binary classification. This method adopts the same approach (user profiling and clustering) as used in PResUP. The hyperparameter details are outlined in Section 6.1.3.

- **1D-CNN based Probing Strategy (CNNPS):** This baseline employs a 1D-CNN with a convolutional layer (10 filters, kernel size 3) followed by max-pooling (pool size: 2). This is followed by flattening and two fully connected dense layers with ReLU activation, finally ending with a sigmoid output layer for binary classification. This method follows the same user profiling and clustering approach utilized in PResUP. The hyperparameter details are presented in Section 6.1.3.

6.1.3 Hyperparameters. To select the optimal values of the hyperparameters for the p-LSTM model, we performed grid search. We tried with the following: (i) batch sizes (8, 16, 24), (ii) epochs (25, 35, 45, 55) (iii) number of p-LSTM nodes (10, 20, 50, 100, 200) and (iv) number of dense layer nodes (10, 20, 50, 100, 200) to train the models. We observed that for the batch size of 16, the epoch of 35, number of p-LSTM nodes of 10, number of dense nodes of 100 (in first dense layer) and number of dense nodes of 50 (in second dense layer), the best classification performance is obtained. Hence, we fixed these values as the model hyperparameters.

For the neural network based baselines (RNNPS, GRUPS, CNNPS), optimal performance was achieved with a batch size of 16, epoch of 35 epochs, and a number of dense nodes of 200. Hence, we fixed these values as the baseline hyperparameters.

6.2 Probing Rate Reduction

We present the comparison of probing rate between PResUP and baselines in Table 4. PResUP has the best probing rate (on average 5.48 probes (std. dev: 1.73) per video per user), while time-based strategy **(TBS)** has the worst (on average 29.89 probes). **TBS** performs the worst as it continuously probes at every 5-second. Similarly, **RePS** also has a very high probing rate (on average 28.79 probes) as it probes based on a rule. **RPS** also returns a high probing rate (on average 17.93) as it probes randomly. Similarly, **PPS** has a high probing rate (on average 21.59) because it relies only on personal self-reports to create the opportune moment detection model. Although the demographic-based baselines have a relatively lower probing rate (**APS**: 7.18, **GBPS**: 6.68), but they still have a higher probing rate than that of PResUP. The **FBS** and **GPS** baslines have comparatively lower probing rate (**FBS**: 5.60, **GPS**: 7.79). As these baselines share data from other users, they reduce the individual user's probes, but could not outperform PResUP. The recurrent neural network-based baseline (**RNNPS**: 6.08) returns lower probing rates than the traditional ML model but not lower than PResUP. The **CNNPS** and **GRUPS** have a probing rate of 6.13 and 7.65, respectively, still inferior to PResUP.

In summary, we note that PResUP reduces the probing rate in comparison to all the baselines (maximum reduction of 81.67% wrt **TBS**, minimum reduction of 2.14% wrt **FBS**, and an average reduction of 34.80%); but whether this reduction in probing rate influences the probing moment detection performance (or the emotion annotation quality) is investigated next.

6.3 Opportune Probing Moment Detection

We next evaluate the probing moment detection performance of PResUP. First, we compare the model performance using the likelihood ratio (LR+) in Table 4. Intuitively, LR+ ($= \frac{TPR}{FPR}$) should be high enough to instill confidence in detecting the positive instances, in our case, opportune moments. LR+ can go as low as 0 meaning if the test is positive, the condition of opportune moment is definitely absent, and as high as infinity (a very large real number, in practice) implying that if the test is positive, the condition of opportune moment is definitely present. Therefore, the model with higher LR+ is desired. We observe that PResUP outperforms (LR+ = 8.61) all the baselines and **TBS** performs the worst (LR+ = 1). Next, we delve deep to investigate if PResUP also has a high TPR and low FPR.

	Probing rate↓	TPR (%) ↑	FPR (%) ↓	LR+ ↑
TBS	29.89 (0.00)	100.00 (0.00)	100.00 (0.00)	1.00
RPS	17.93 (0.01)	60.28 (0.01)	59.98 (0.01)	1.00
RePS	28.79 (0.01)	99.39 (0.02)	95.77 (0.01)	1.04
PPS	21.59 (3.46)	71.53 (0.23)	9.96 (0.04)	7.18
FBS	5.60 (4.58)	53.23 (0.03)	9.53 (0.17)	5.59
APS	7.18 (5.32)	65.27 (0.24)	15.88 (0.16)	4.11
GBPS	6.68 (4.96)	66.84 (0.24)	14.32 (0.14)	4.67
GPS	7.79 (3.11)	76.21 (0.18)	19.72 (0.11)	3.86
RNNPS	6.08 (3.25)	61.16 (0.38)	14.08 (0.12)	4.34
GRUPS	7.65 (3.60)	67.50 (0.36)	18.83 (0.09)	3.58
CNNPS	6.13 (1.75)	68.80 (0.19)	13.13 (0.08)	5.24
PResUP	**5.48 (1.73)**	**80.07 (0.16)**	**9.30 (0.05)**	**8.61**

Table 4. Performance comparison of PResUP and baselines with respect to different metrics. The values indicate the average for a metric across all users. The values in the parenthesis indicate std. dev. PResUP outperforms all baselines in terms of probing rate (having the least probing rate, thereby reducing annotation effort the most). It also outperforms all baselines in terms of LR+ (having the highest LR+) and FPR (having the least FPR). Although **TBS** and **RePS** have the highest and second highest TPR, they also have the worst (100%) and second worst FPR (95.77%) respectively. ↑, ↓ indicate higher and lower value preferred respectively.

6.3.1 TPR Analysis. We present the comparison of mean TPR in detecting the opportune probing moments in Table 4. The **FBS** baseline performs the worst (mean TPR: 53.23%, std dev: 0.03%) followed by the **RPS** baseline (mean TPR: 60.28%, std dev: 0.01%). The **RPS** model does not perform well as it randomly probes at some window, and developing a feature-based model like **FBS** without considering the temporal sequence is also not a good option. The **PPS** (mean TPR: 71.53%, std dev: 0.23%) highlights that relying only on the personal data to train the opportune moment detection model (as done in the **PPS**) does not yield good performance. However, the sequence-based model **(GPS)** that aggregates data from all users (without considering the physiological response similarity) returns comparatively better performance (mean TPR: 76.21%, std dev: 0.18%). Similarly, the demographic-based baselines (**APS, GBPS**) also return relatively better performance with mean TPR of 65.27% (std dev: 0.24%) and 66.84% (std dev: 0.24%) respectively. We also note that the neural network-based other baseline models (**RNNPS**: 61.16% (std dev: 0.38%), **GRUPS**: 67.50% (std dev: 0.36%), **CNNPS**: 68.80% (std dev: 0.19%)) do not perform well in terms of TPR. But PResUP outperforms all these baselines with a mean TPR of 80.07% (std dev: 0.16%).

These findings demonstrate that - (a) aggregating data from all users without the physiological response similarity (as done in **GPS**), or (b) combining data just based on demographic similarity (as done in **GBPS,** and **APS**) are not good choices. Rather, aggregating data from users having similar physiological responses (as done in PResUP) helps to obtain superior performance. Finally, we note that **TBS** has the best mean TPR (100%). This is because it continuously probes at every 5-seconds interval, therefore all opportune probing moments are captured but at the cost of very high probing rate (Section 6.2) and very high error rate (Section 6.3.2). Another baseline **RePS** also has very high TPR (99.39%). However, this is not useful as it has a very high probing rate (Section 6.2) and a significantly high error rate (Section 6.3.2). In summary, these results demonstrate that leveraging the user's physiological response profile similarity and a sequence of previous time steps (as done in PResUP) improve opportune probing moment detection performance.

6.3.2 Error Analysis. In this section, we analyze the error rate in detecting the opportune probing moments (see Table 4). We note that PResUP outperforms all the baselines with a mean FPR of 9.30%. The **TBS** model

performs the worst (mean FPR: 100%) as it does the probing continuously and therefore ends up probing in every inopportune moment. The **RePS** has the second worst performance (mean FPR: 95.77%) as it probes based on a rule. Also, **RPS** model's performance is not good (mean FPR: 59.98%) as it randomly probes in some windows for annotation collection. The **PPS, FBS,** and **GPS** baselines also have a high mean FPR of 9.96%, 9.53%, and 19.72%, respectively. The demographic based models (**APS, GBPS**) also have a high mean of FPR of 15.88%, and 14.32%, respectively. Similarly, the neural network-based models (**RNNPS**: 14.08%, **GRUPS**: 18.83%, **CNNPS**: 13.13%) also have high mean FPR. These findings further underscore the effectiveness of PResUP - that it not only detects the opportune moments correctly (as TPR is high), but also makes relatively few errors while detecting the opportune probing moments.

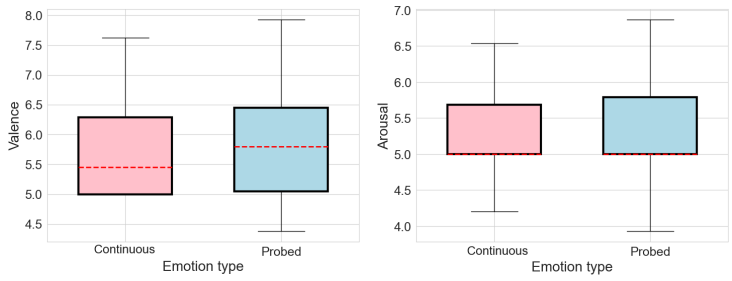
6.4 System Performance: Latency Analysis of PResUP

In this section, we analysed the performance of PResUP in terms of latency. In specific, we measured the time taken in each phases of the framework. We carried out this measurement on a computer with following configuration - a 13th Gen Intel(R) Core(TM) i9-13900 processor at 2.00 GHz, 128 GB of RAM (128 GB usable), and a 64-bit operating system (Windows 11) with an x64-based processor.

We ran all the phases (profile creation, similar user clustering, and opportune moment detection; as outlined in Section 5) 15 times on the data collected from the user study I and computed the time required at each phase. We noted that the time required profile creation is 66.70 milliseconds on average (SD: 3.31), clustering is 99.75 milliseconds on average (SD: 2.61), and opportune moment prediction is 0.62 milliseconds on average (SD: 0.17). Notably, the time required for opportune moment prediction is very low, as the prediction time is computed for every segment. These values highlight that that the framework does not have a very high latency.

6.5 Emotion Annotation Quality Comparison

We next investigate the annotation quality of valence and arousal. In specific, we compare the user-wise and video-wise valence (and arousal) sampled using PResUP (at opportune moments) and present in original continuous annotations. We observe that in both the cases (user-wise and video-wise), there is no significant difference in the sampled and continuous data of valence (and arousal). We present the user-wise comparison next. However, to improve readability, the detailed video-wise comparison is presented in Appendix A.3.



(a) Comparison of valence scores **(b)** Comparison of arousal scores

Fig. 7. Comparison of users' valence and arousal between the ground truth (continuous annotation) and probed moments' annotation (using PResUP) reveals no significant difference between the continuous and probed values. Mann-Whitney U test shows both valence and arousal do not vary significantly ($p < 0.05$) between the continuous and probed annotations.

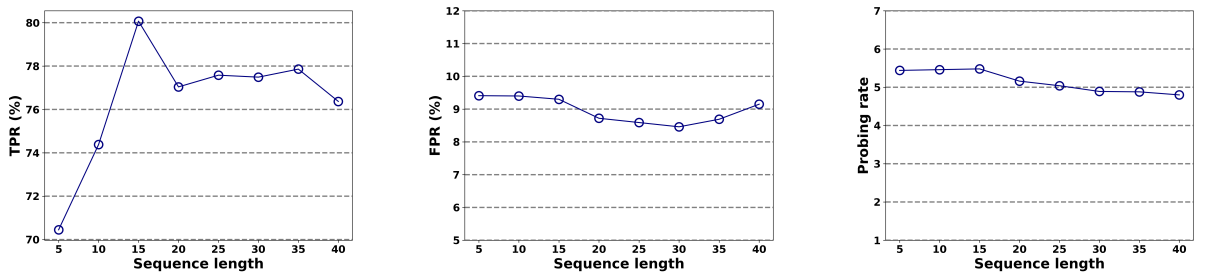
To compare the users' valence (and arousal), we perform the following steps (once for valence and once for arousal). We compute the median score of valence for every user by collecting annotations only at those

windows, when the probe was issued as per the PResUP framework. Similarly, we compute the user-wise median score of valence from the ground truth data. Then, we checked if the valence (continuous and sampled) and arousal (continuous and sampled) follow normal distribution using the Shapiro-Wilk test [54]. This experiment revealed that the responses did not follow a normal distribution ($p < 0.05$) both for valence and arousal. Since the distribution is not normal and the two groups (continuous, sampled) are not paired, we performed Mann-Whitney U test [41] to evaluate the difference between continuous and sampled valence scores. The same steps were repeated for arousal.

The comparison for the valence and arousal are shown in Fig. 7a, and 7b, respectively. The medians of valence (continuous) and valence (probed) are 5.45 and 5.8, respectively. The Mann-Whitney's U test did not find a significant difference in the continuous and probed values ($U = 754.0$, $Z = 1.21$, $p = 0.229$) of valence. Similarly, the medians of arousal (continuous) and arousal (probed) are 5.0 (in both cases). In arousal also, we did not find a significant difference in the continuous and probed values ($U = 603.5$, $Z = 0.520$, $p = 0.607$) from the Mann-Whitney's U test. In summary, these findings underscore that there is no significant difference in the valence and arousal values between the ground truth continuous annotations and the values sampled using the PResUP framework.

6.6 Influence of Temporal Sequence

In this section, we investigate the influence of temporal sequence on the performance of PResUP. Since the p-LSTM model of PResUP takes a number of previous segments as input to determine whether the current segment is opportune (or not), it is required to identify the optimal sequence length.



(a) TPR across sequence length **(b)** FPR across sequence length **(c)** Probing rate across sequence length

Fig. 8. Influence of sequence length on (a) TPR, (b) FPR, and (c) Probing rate. The highest TPR is obtained for sequence length of 15. At the same time, if sequence length is increased beyond 15, although the probing rate and FPR improve, the improvement is not significant (in comparison to the deterioration of TPR).

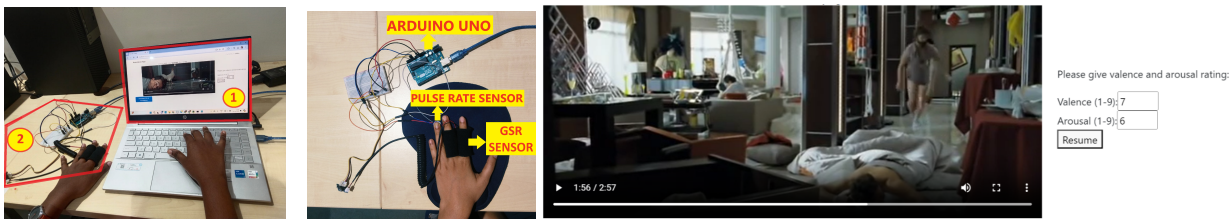
To investigate the optimal sequence length, we vary the number of sequence length from 5 to 40 as input to the the p-LSTM network. We choose 40 as the highest sequence length, as the longest video has 40 segments. We plot the variation in TPR, FPR and probing rate in Fig. 8. We observe that with increasing segments, initially the TPR increases (till sequence length 15) and then it stabilizes after dropping (beyond 15). We obtained the highest TPR (80.07%) for a sequence length of 15 (Fig. 8a). Moreover, we do not observe a large variation in the FPR with increasing sequence length (Fig. 8b). So, we decided to use 15 as the sequence length in our experimental results (presented earlier in Section 6). The variation in TPR with sequence length can be attributed to the persistence effect of emotion [63], which suggest that the human emotion persists for some time once experienced. Typically, many emotions have a persistence period of a few seconds to minutes [62]. Therefore, if the sequences are very

long (i.e., beyond the persistence period of an emotion), the earlier emotion may not have an effect on the current segment's emotion. As a result, trying to estimate the current segment as opportune (or not) based on a number of segments way behind in time may not yield a good detection performance.

We also note that the probing rate drops gradually with an increasing number of sequences (Fig. 8c). This can be attributed to the fact as the number of segments are increased, the number of queries sent to the model is reduced (which in turn reduces the number of probes to be issued). For example, for a sequence length of 20, the first query to the model will be sent only when 20 segments are accumulated, whereas for a sequence length of 10, in the same period 11 (20-10+1) queries would be made. Moreover, there is not a substantial decrease in the probing rate with increasing sequence length. As a result, considering the values of TPR, FPR, and probing rate, the sequence length (=15) was selected.

7 User Study II: Deployment of PResUP

To investigate the effectiveness of PResUP for an unknown user in a real-world setting, we deployed it and performed the second user study. Unlike the previous user study (Section 3), in this study we integrated the ML model for opportune moment detection with the annotation application and probe user for self-report only at the opportune moments (as detected by the ML model). In this section, we discuss the user study in detail, including the experiment apparatus, study procedure, and the dataset. We obtained the IRB approval (HEC/BPGC/2023/005) from our institute's Human Ethics Committee (HEC) prior to the user study.



(a) User study setup **(b)** Sensor setup **(c)** "Annotate App" (modified UI with emotion collection pop-up)

Fig. 9. The overview of user study after deployment of PResUP- (a) study setup (including the sensors and the UI for video consumption and annotation opportunistically) (b) Sensor configuration for physiological response (GSR, HR) collection. (c) "Annotate App" modified UI for video consumption and annotation. The interface displays the video and allows the user to record emotion (valence, arousal on a scale of 1 to 9) at the opportune moments by pausing the video. Once the annotation is provided and the user clicks on the 'Resume' button on the self-reporting pop-up, the self-reporting pop-up disappears and the video resumes to play.

7.1 Experiment Apparatus and Deployment Scenario

We show the complete user study setup in Fig. 9. We used the same experiment apparatus (Fig. 9a) as introduced in Section 3.1. It also consists of the two components - the sensor setup (Fig. 9b) and the annotation application ("Annotate app", (Fig. 9c)). We do not make any changes to the sensor setup (comprising a GSR sensor and pulse rate sensor), which the user wears during the study. However, we make the following modifications in the annotation application.

First, we changed the interface of the annotation application so that it captures the annotations *only* at opportune moments (instead of continuous annotation). Therefore, the new UI does not have the provision for capturing continuous annotation, rather it probes (by sending a popup) the user at opportune (as determined by the model) moments during video consumption. The annotation interface and the emotion self-report collection

pop-up are shown in Fig 9c. Once the pop-up is displayed, the video pauses and the user records the valence and arousal (on a scale of 1 to 9) by typing the corresponding key from the keyboard. After providing ratings, once the user presses the 'Resume' button, the corresponding video segment is labeled with the values as provided by the user, the emotion self-reporting pop-up disappears and the video resumes. As the user watches the video, we passively collect the HR and GSR sensor values, however, when the video is paused, we do not record the sensor values.

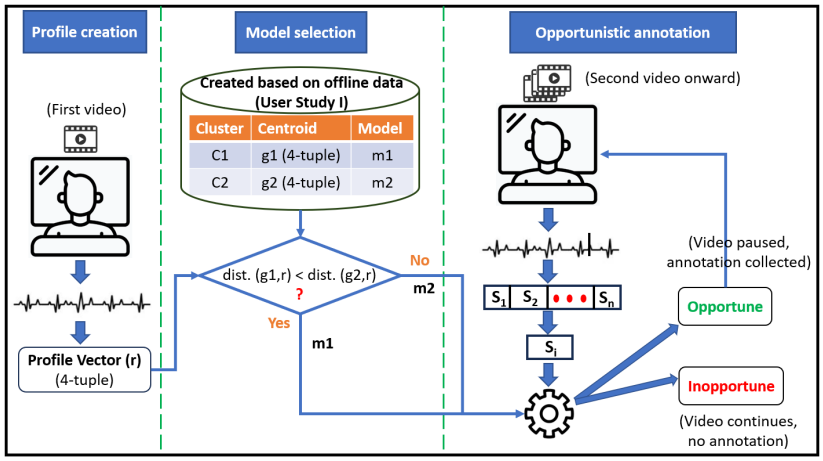


Fig. 10. Deployment of the PResUP framework. In the deployment scenario, first, we create the physiological profile of the (unknown) user based on the first video. Next, this profile vector is passed to the database (which contains the cluster details based on the offline data collected earlier - Section 5.2) to find the nearest cluster and the corresponding opportune moment detection model. Once the model is selected, it is used for detecting the opportune moments based on the physiological response signal as recorded while watching the subsequent videos (second video onward). If the current segment is detected as inopportune, the video continues and no probing is done; otherwise the video is paused and the user is probed for the emotion self-report.

Second, we also integrated the opportune moment detection models with the application. We maintained a record of the clusters and models created earlier using the data collected from the first user study (N=36, Section 3). Notably, there were two clusters (and corresponding models, Section 5.2, 5.3) created from the first user study. When an unknown user (say u) starts using the application, we sample the HR and GSR from the sensor setup, but do not probe the user for emotion self-report while the user watches the first video. Rather, the sensor data collected during this period is used to create the profile vector (Section 5.1) of the user in real-time. Once the vector is created, we compute the distance of this user's profile vector from the centroids of the two clusters and identify the user's (u) nearest cluster. Accordingly, we enable the opportune moment detection model associated with the nearest cluster for opportune moment detection. Thereafter, for all videos (second video onward), we segment the sensor data (into 5-sec window) and send to the model to decide if the current segment is opportune for probing. If the current segment is opportune, we pause the video, and probe the user issuing the self-report collection pop-up (Fig. 9c), otherwise, we do not probe the user (and the video continues). Once, the user provides the emotion self-report on the pop-up and clicks on the 'Resume' button, the pop-up disappears and the video resumes. This process continues until the user watches all the videos. We show the schematic diagram for the deployment scenario in Fig. 10.

7.2 Study Procedure

We recruited 18 participants (5F, 13M) aged between 20 and 30 years from our university. Notably, these participants differ from those who participated in the first user study (Section 3). During the study, each participant watched the same eight videos (as used earlier, Table 1) and recorded the emotion annotations opportunistically (as determined by the opportune moment detection model integrated in the experiment apparatus). The participants were explained about valence and arousal. They were also instructed to record integer values between 1 and 9.

Similar to the first user study, every participant watched the videos in a predetermined random order and recorded the valence and arousal scores on a scale of 1 to 9 whenever the self-report collection pop-up appeared. A two-minute-long blue screen interleaved the videos to avoid the carry-over effect.

7.3 Final Dataset

We obtained a total of 4524 segments (5-second window) from the 18 users. The users watched the 8 videos in a predetermined random order. For every user, the first video was used to create the profile vector and during this period no probing was performed. As a result, the segments associated in the first videos (a total of 612 segments, $\approx$ 34 segments per user) were not considered for probing. The remaining segments (a total of 3912 segments, $\approx$ 217 segments per user) associated with the other seven videos were considered for probing (i.e., probed if detected opportune by the model). Out of these 3912 segments, 703 segments were predicted as opportune by the model and in these segments probing were done. This number yields an average probing rate of 5.58 (per user per video).

8 Evaluation: User Study II

In this section, we evaluate the performance of PResUP after deployment. To evaluate, we compare the performance of PResUP with the baselines (Section 6.1.2) using the same performance metrics (Section 6.1.1). Additionally, we performed a post-study qualitative evaluation of PResUP.

8.1 Data Labeling

To validate the performance of PResUP, we are required to know the label (ground truth) of each segment. We adopted the same data labeling steps (segmentation, opportune segment labeling) as outlined in Section 3.3 to tag a segment as opportune (or not). In total, we marked 16.8% segments as opportune and 83.2% segments as inopportune. Notably, this distribution of opportune and inopportune segments are in line with the data collected in the first user study (Section 3.4). These labels are used as the ground truth and the predicted values are decided based on the model output (i.e., if a user is probed at a specific segment of the video, then the segment is predicted as opportune, and the vice-versa). We compare these two to find the values of the different metrics.

8.2 Quantitative Evaluation

In this section, we compare the performance of the PResUP framework with the baselines. We provide the comparison in Table 5. We discuss the performance comparison with respect to the metrics next.

8.2.1 Probing Rate. The results demonstrate that PResUP outperforms all the baselines with a probing rate of 5.79 (std. dev: 2.54). This probing rate yields a reduction of 38.05% in terms of the average probing rate of the baselines (max reduction: 83.18% with respect to the **TBS** baseline, min reduction: 0.34% with respect to the **FBS** baseline). These findings are similar to the offline analysis (Section 6.2), which indicates that PResUP reduces the probing rate substantially when deployed for opportunistic annotation.

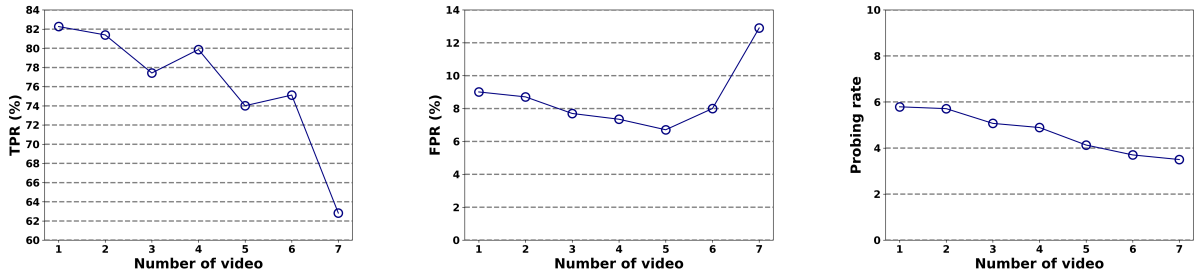
	Probing rate↓	TPR (%) ↑	FPR (%) ↓	LR+ ↑
TBS	34.42 (0.00)	100 (0.00)	100 (0.00)	1.00
RPS	23.99 (0.01)	69.79 (0.01)	69.6 (0.01)	1.00
RePS	30.38 (0.10)	98.82 (0.01)	96.25 (0.00)	1.03
PPS	12.16 (1.62)	78.8 (0.18)	9.49 (0.03)	8.30
FBS	5.81 (2.57)	61.62 (0.15)	9.21 (0.02)	6.69
APS	7.77 (5.21)	71.23 (0.20)	12.11 (0.12)	5.88
GBPS	7.65 (4.94)	73.82 (0.20)	11.3 (0.11)	6.53
GPS	8.22 (4.02)	64.77 (0.41)	17.54 (0.18)	3.69
RNNPS	8.09 (3.94)	54.9 (0.36)	21.29 (0.10)	2.58
GRUPS	7.36 (3.77)	51.16 (0.39)	19.7 (0.09)	2.60
CNNPS	9.59 (1.98)	34.82 (0.18)	31.13 (0.06)	1.12
PResUP	**5.79 (2.54)**	**82.26 (0.09)**	**9.01 (0.06)**	**9.13**

Table 5. Performance comparison of PResUP (after deployment) and baselines with respect to different metrics. The values indicate the average for a metric across all users. The values in the parenthesis indicate std. dev. PResUP outperforms all baselines in terms of probing rate (having the least probing rate, thereby reducing annotation effort the most). It also outperforms all baselines in terms of LR+ (having the highest LR+) and FPR (having the least FPR). Although **TBS** and **RePS** have the highest and second highest TPR, they also have the worst (100%) and second worst FPR (96.25%) respectively. ↑, ↓ indicate higher and lower value preferred respectively.

8.2.2 Opportune Moment Detection Performance. We compare the opportune moment detection performance in terms of the following metrics - LR+, TPR, and FPR (see Table 5). PResUP outperforms all the baselines with the highest LR+ of 9.13. We also investigate whether PResUP detects the opportune moments accurately and generates a few false probes, i.e., it has a high TPR and low FPR. PResUP also has the highest TPR (ignoring the **TBS** and **RePS** baselines as both of them also have very high FPR) of 82.26% and the lowest FPR (9.01%). Notably, these values are similar to the earlier results (Table 4) carried out during offline analysis. In summary, these findings indicate that PResUP, when deployed, reduces the probing rate substantially and yet detects the opportune moments correctly.

8.2.3 Influence of Number of Videos used for Profile Creation. In this section, we evaluate the performance of PResUP in the deployment scenario based on the number of videos used for profile creation. This is important to investigate as it allows to find out the optimal number of videos required for profile creation. To find out the optimal number of videos required for user profile creation, we vary the number of videos used for profile creation from 1 to 7 and measure the performance of PResUP in terms of TPR, FPR, and probing rate (Fig. 11). Notably, if 'n' videos are used for profile creation, the remaining (8-n) videos are used for performance assessment.

We observe that the TPR reduces with an increasing number of videos being used for profile creation (Fig. 11a). This may be attributed to the fact that with more videos, the user profile contains a lot of variance, which may lead to incorrect cluster identification and poor performance. We also observe in Fig. 11b that FPR reduces a little with the increasing number of videos; then again, it increases (beyond 5). However, the total variation noted in FPR is not substantial if the number of videos is increased from 1 to 7. The probing rate reduces gradually if more videos are used for profile creation (Fig. 11c). This is because if more videos are used for profile creation, the model is therefore probed for fewer segments to detect the opportune segments, which in turn reduces the probing rate. Notably, although the probing rate reduces with an increasing number of videos, the reduction is not substantial. On the contrary, the TPR drops sharply if more videos are used for profile creation. Therefore, we decide to use only one video for user profile creation during deployment (Section 7.1).



(a) TPR across number of videos used for profile creation

(b) FPR across number of videos used for profile creation

(c) Probing rate across number of videos used for profile creation

Fig. 11. Influence of the number of videos used for profile creation on (a) TPR, (b) FPR, and (c) Probing rate. The highest TPR is obtained if one video is used for profile creation. At the same time, if more videos are used for profile creation, although the probing rate and FPR improve, the improvement is not significant (in comparison to the deterioration of TPR).

8.2.4 System Performance: Latency Analysis of PResUP after Deployment. In this section, we measure the latency of PResUP after deployment. We used the same computer configuration (as mentioned in Section 6.4) to measure the latency values of each component of the framework. Notably, we ran the framework for each of the users (N=18) and computed the average time required in each phase. In the deployment scenario, the profile is created only when the user completes watching the first video. Therefore, the profile creation time includes the duration of the first video. As a result, the average time required for profile creation is 158055.38 milliseconds (SD: 21.42). As evident, this is relatively higher as the profile creation step needs to wait (and hence this is not a system limitation) for the first video to complete. The time required for clustering is 67.66 milliseconds on average (SD: 6.14), and opportune moment prediction is 0.87 milliseconds on average (SD: 0.21). These values demonstrate that the end-to-end time required in the framework after deployment (for opportune moment detection) is on average 158122.9 milliseconds, which is not very high.

8.3 Post-study User Survey

We also performed a qualitative evaluation to understand the usability of the PResUP framework. In specific, we conducted a post-study user survey to understand - the timeliness of probing (i.e., probing was done at the opportune moment), interruption level, ease of using the interface, and user experience.

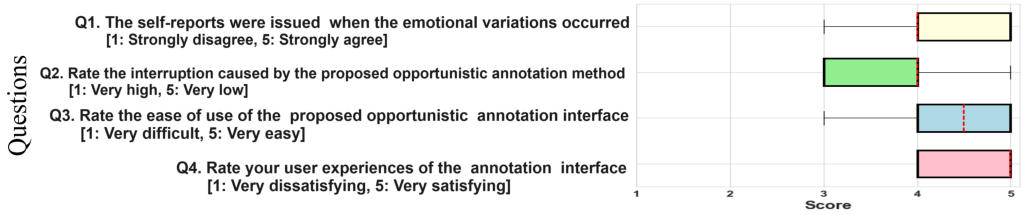


Fig. 12. Boxplot showing the scores as obtained from the survey participants for the four survey questions in the questionnaire. Higher values for each of the four questions are desired.

We performed a qualitative survey (immediately after the user study II) asking user feedback (on a scale of 1 to 5) for the following questions - (a) the self-report probes were issued when the user's emotion changed [1: strongly disagree to 5: strongly agree] (b) the level of interruption caused by the probing [1: very high to 5: very low interruption] (c) the easiness of using the interface [1: very difficult to 5: very easy] (d) the user experience for user interface [1: very dissatisfying to 5: very satisfying].

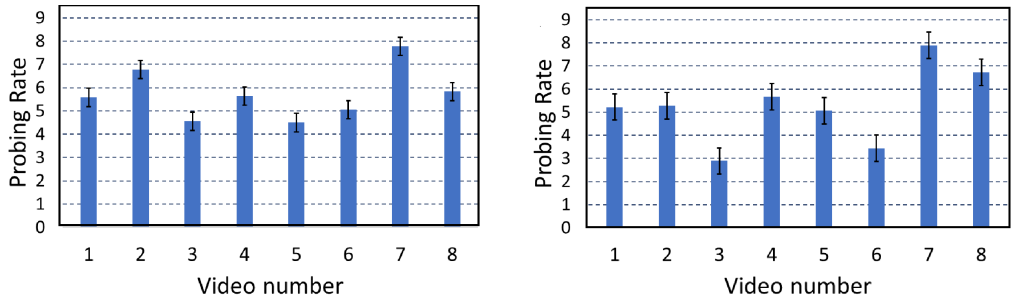
We present the findings from this survey in Fig. 12. We plotted the scores against each question as obtained from all the 18 users, who participated in this study. We observe that most of the users agree that the probes were issued when the emotions actually varied (median is 4 and upper quartile is 5). We also noted that the users feel the interruption caused by the interface is low (median is 4). However, a comparatively low upper quartile value (of 4) indicates that the interruption level needs to be further improved. The users agreed that the interface is easy to use (median is 4.5 and upper quartile is 5) and they provided a high user satisfaction score (median and upper quartile; both are 5). These findings as obtained from the qualitative evaluation highlight the usability of the PResUP framework in terms of collecting user annotation based on user emotion variation with low interruption and high user satisfaction.

9 Discussion

In this section, we discuss the implications of the findings. We also mention the possible future works in this direction.

9.1 Implications of the Findings

9.1.1 Reducing Continuous Annotation Effort: The major takeaway from this paper is that PResUP reduces the continuous emotion annotation overhead during video consumption without compromising the annotation quality. We investigated deep to find out if the probes issued vary across the videos in different user studies. We show the average number of probes (per user) issued for every video for user study I (see Fig. 13a) and user study II (see Fig. 13b) in Fig. 13. The analysis reveals that in both the studies none of the videos has a very high number of probes. The comparatively high number of probes issued for video 7 and 8 can be attributed to the fact that these videos embed the 'scary' emotion, which may have caused a high number of emotional variation (and therefore, a higher number of probes). We also note that in both the studies, a similar number of probes were issued in each video (barring video 2, 3). These could be attributed to the different groups of subjects, who participated in two studies. In summary, we do not observe a large number of probes issued in any video for both the studies.



(a) Video-wise avg. probes per user (user study I) **(b)** Video-wise avg. probes per user (user study II)

Fig. 13. The average number of probes (per user) issued in every video (a) for user study I, (b) for user study II reveals that there is not a large number of probes issued in any videos

Further, we observe that PResUP is efficient when deployed, implying that it can reduce the probing rate and detect the opportune probing moments accurately for an unknown user (Section 8). These findings suggest that it is possible to develop opportunistic, low-overhead emotion annotation strategy and yet collect high-quality emotion annotations. We envision that PResUP can be generalized to other modality also (e.g., audio) that reveal variations in physiological responses.

9.1.2 Physiological Response based User Profile: Another key takeaway from the current study is the physiological response based user profile creation approach. This approach allows to create a snapshot of user behavior with respect to different physiological signals. Notably, this profile can be used for user modeling, personalization, and recommendation. We have shown the possibility of aggregating data from similar users (in terms of user profiles). The same approach can be adopted in other domains (e.g., while recommending some content based on user's behavioral response). We have also shown that finding similar users just based on the demographic profile is not optimal, rather adopting the approach to cluster users based on profile similarity helps to obtain superior results (Table 4, 5).

9.1.3 Generalizability of PResUP:. We evaluated the generalizability of PResUP on two publicly available continuous emotion annotations dataset (CASE [56], K-emocon [47]). The CASE dataset contains the physiological responses and continuous emotions annotations (valence and arousal) of 30 users as they watch eight different videos. The K-emocon dataset also contains physiological responses and continuous annotations of valence and arousal from 16 paired debate sessions. We applied PResUP on these two datasets by adopting the steps (segmentation, user clustering, and opportune moment detection modeling) and compared the performance with the baselines (Section 6.1.2). In both the datasets, we observe that PResUP outperforms all the baselines in terms of probing rate, TPR and FPR. In case of CASE dataset, we obtain an average reduction of 26.19% in the probing rate, whereas in case of K-emocon the average probing rate reduction is 30.47%. We present the detailed comparison results in Appendix A (CASE daataset: Section A.4, K-emocon dataset: Section A.5). These findings underscore the generalizability of PResUP.

9.1.4 Deployment Consideration for Potential Use Case. We deployed PResUP for continuous emotion annotation reduction and demonstrated its effectiveness (Section 7). However, we envision that the framework can be deployed in different use cases with some scenario-dependant adaptations. For example, PResUP can be deployed to monitor the continuous stress of students during online video lectures. The framework can detect the moments of high stress and accordingly adapt the lecture delivery. To deploy for such a use case, the framework may need minor modifications (e.g., inclusion of video lectures instead of the stimuli videos, incorporating the stress detection model based on the physiological signals, defining the high stress moments as opportune moments to adapt the lecture delivery, creating user profiles based on high stress and low stress moments). In summary, by adapting to the specific context, the proposed framework can be used in different situations.

9.2 Future Works

In this section, we discuss the future works and the limitations of the current framework. First, we aim to investigate the relationship between probing rate and opportune moment detection performance so that depending on the priority (e.g., to have minimum user interruption measured by least probing rate, to accurately detect all the annotation points measured by highest TPR) we can select the best model for a user. Second, in this work, we did not consider other physiological signals (e.g., respiration or skin temperature) for opportune moment detection. Considering those signals may improve the opportune moment detection performance as these signals also correlate with human emotion [56]. Third, as users are spending more time on mobile devices for consuming the videos, we aim to find the applicability of PResUP on a mobile platform. We also aim to reduce the interruption

level of the current framework (by reducing the FPR) so that the users' viewing experience is further improved. Finally, automatic identification of sufficient physiological data for profile creation also remains to be in the scope of future work.

10 Conclusion

This paper proposes PResUP, a framework for opportunistic emotion self-report collection to reduce the continuous emotion annotation effort during video consumption. The key idea is that physiological signal variations can be leveraged for opportunistic emotion self-report collection to alleviate the need for continuous annotation. To implement this, first, the framework constructs the physiological response profile of the users and clusters the users based on the profile similarity. Later, combining physiological response data among the users in a cluster, a parameterized LSTM (p-LSTM) model is constructed to detect the opportune probing moments for emotion self-report collection. The profile creation and data sharing among similar users allow to train the p-LSTM model with a larger dataset, thereby improving the opportune moment detection performance. We validated the proposed approach conducting a real-world user study (N=36). The major findings from the evaluation are that PResUP reduces the probing rate by 34.80% (on average), detects the opportune probing rate with a TPR of 80.07% (on average), and yet maintains similar emotion scores as present in the continuous annotations. We also deployed the proposed framework by running a follow-up user study (N=18). The findings from this study also align with the earlier findings (average probing rate reduction: 38.05%, average TPR: 82.26%). In summary, PResUP allows to reduce the continuous emotion annotation effort during video consumption leveraging the physiological response variation.

11 Acknowledgments

This research has been supported by the Chanakya Ph.D. Fellowship at AI4ICPS Innovation Hub (IIT Kharagpur), the SURE grant (SUR/2022/001965) of SERB (Science and Engineering Research Board) of the Department of Science & Technology (DST), Government of India, and the CDRF grant (C1/23/152) of BITS Pilani Goa. S.S would like to thank the CDRF grant (C1/23/114) of BITS Pilani K K Birla Goa Campus and SERB CRG- DST, GoI (CRG/2023/003210) for supporting the work.

References

[1] Akhilesh Adithya, Snigdha Tiwari, Sougata Sen, Sandip Chakraborty, and Surjya Ghosh. 2022. OCEAN: Towards Developing an Opportunistic Continuous Emotion Annotation Framework. In *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*. IEEE, 9–12.

[2] Foteini Agrafioti, Dimitris Hatzinakos, and Adam K Anderson. 2011. ECG pattern analysis for emotion detection. *IEEE Transactions on affective computing* 3, 1 (2011), 102–115.

[3] Zeeshan Ahmad and Naimul Khan. 2022. A Survey on Physiological Signal-Based Emotion Recognition. *Bioengineering* 9, 11 (2022), 688.

[4] Khaled Alrawashdeh and Carla Purdy. 2018. Fast activation function approach for deep learning based online anomaly intrusion detection. In *2018 IEEE 4th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing,(HPSC) and IEEE International Conference on Intelligent Data and Security (IDS)*. IEEE, 5–13.

[5] Samaneh Aminikhanghahi and Diane J Cook. 2017. A survey of methods for time series change point detection. *Knowledge and information systems* 51, 2 (2017), 339–367.

[6] Rahimpour Cami Bagher, Hamid Hassanpour, and Hoda Mashayekhi. 2017. User trends modeling for a content-based recommender system. *Expert Systems with Applications* 87 (2017), 209–219.

[7] Patricia Bota, Pablo Cesar, Ana Fred, and Hugo Placido da Silva. 2024. Exploring Retrospective Annotation in Long-videos for Emotion Recognition. *IEEE Transactions on Affective Computing* (2024).

[8] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.

[9] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[10] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42

(2008), 335–359.
[11] Taryn Chalmers, Blake Anthony Hickey, Phillip Newton, Chin-Teng Lin, David Sibbritt, Craig S McLachlan, Roderick Clifton-Bligh, John Morley, and Sara Lal. 2021. Stress watch: The use of heart rate and heart rate variability to detect stress: A pilot study using smart watch wearables. *Sensors* 22, 1 (2021), 151.
[12] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.
[13] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20 (1995), 273–297.
[14] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey, and Marc Schröder. 2000. 'FEELTRACE': An instrument for recording perceived emotion in real time. In *ITRW Speech-Emotion*.
[15] Roddy Cowie, Martin Sawey, Cian Doherty, Javier Jaimovich, Cavan Fyans, and Paul Stapleton. 2013. Gtrace: General trace program compatible with emotionml. In *2013 humaine association conference on affective computing and intelligent interaction*. IEEE, 709–710.
[16] Sandra De Amo, Mouhamadou Saliou Diallo, Cheikh Talibouya Diop, Arnaud Giacometti, Dominique Li, and Arnaud Soulet. 2015. Contextual preference mining for user profile construction. *Information Systems* 49 (2015), 182–199.
[17] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2018. Unobtrusive assessment of students' emotional engagement during lectures using electrodermal activity sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–21.
[18] Sidney K D'mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)* 47, 3 (2015), 1–36.
[19] Maria Egger, Matthias Ley, and Sten Hanke. 2019. Emotion recognition from physiological signal analysis: A review. *Electronic Notes in Theoretical Computer Science* 343 (2019), 35–55.
[20] David L Elliott. 1993. *A better activation function for artificial neural networks*. University of Maryland. Systems Research Center.
[21] Hans Eysenck. 2018. *Dimensions of personality*. Routledge.
[22] Karl Friston, Rosalyn Moran, and Anil K Seth. 2013. Analysing connectivity with Granger causality and dynamic causal modelling. *Current opinion in neurobiology* 23, 2 (2013), 172–178.
[23] Qian Gao, Su Mei Xi, and Young Im Cho. 2013. A multi-agent personalized ontology profile based user preference profile construction method. In *IEEE ISR 2013*. IEEE, 1–4.
[24] Yang Gao, Yincheng Jin, Seokmin Choi, Jiyang Li, Junjie Pan, Lin Shu, Chi Zhou, and Zhanpeng Jin. 2021. Sonicface: Tracking facial expressions using a commodity microphone array. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–33.
[25] Hernán F García, Mauricio A Álvarez, and Álvaro Á Orozco. 2016. Gaussian process dynamical models for multimodal affect recognition. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 850–853.
[26] Jeffrey M Girard and Aidan GC Wright. [n. d.]. DARMA: Software for dual axis rating and media annotation. *Behavior research methods* 50, 3 ([n. d.]).
[27] Zied Guendil, Zied Lachiri, Choubeila Maaoui, and Alain Pruski. 2015. Emotion recognition from physiological signals using fusion of wavelet based features. In *2015 7th International Conference on Modelling, Identification and Control (ICMIC)*. IEEE, 1–6.
[28] Ido Guy, Naama Zwerdling, Inbal Ronen, David Carmel, and Erel Uziel. 2010. Social media recommendation based on people and tags. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 194–201.
[29] Satchit Hari, Sayan Sarcar, Sougata Sen, and Surjya Ghosh. 2022. AffectPro: Towards Constructing Affective Profile Combining Smartphone Typing Interaction and Emotion Self-reporting Pattern. In *Proceedings of the 2022 International Conference on Multimodal Interaction*. 217–223.
[30] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28, 1 (1979), 100–108.
[31] Muhammad Asif Hasan, Nurul Fazmidar Mohd Noor, Siti Soraya Binti Abdul Rahman, and Mohammad Mustaneer Rahman. 2020. The transition from intelligent to affective tutoring system: a review and open issues. *IEEE Access* 8 (2020), 204612–204638.
[32] Shenda Hong, Yuxi Zhou, Junyuan Shang, Cao Xiao, and Jimeng Sun. 2020. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers in Biology and Medicine* 122 (2020), 103801.
[33] Yu-Liang Hsu, Jeen-Shing Wang, Wei-Chun Chiang, and Chien-Han Hung. 2017. Automatic ECG-based emotion recognition in music listening. *IEEE Transactions on Affective Computing* 11, 1 (2017), 85–99.
[34] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. 2018. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 387–395.
[35] Md Adnanul Islam, Md Saddam Hossain Mukta, Patrick Olivier, and Md Mahbubur Rahman. 2022. Comprehensive guidelines for emotion annotation. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*. 1–8.
[36] Eiman Kanjo, Eman MG Younis, and Chee Siang Ang. 2019. Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Information Fusion* 49 (2019), 46–56.

[37] Joseph E LeDoux and Stefan G Hofmann. 2018. The subjective experience of emotion: a fearful view. *Current Opinion in Behavioral Sciences* 19 (2018), 67–72.
[38] Shuyu Lin, Ronald Clark, Robert Birke, Sandro Schönborn, Niki Trigoni, and Stephen Roberts. 2020. Anomaly detection for time series using vae-lstm hybrid model. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Ieee, 4322–4326.
[39] Yuan-Pin Lin, Chi-Hong Wang, Tzyy-Ping Jung, Tien-Lin Wu, Shyh-Kang Jeng, Jeng-Ren Duann, and Jyh-Horng Chen. 2010. EEG-based emotion recognition in music listening. *IEEE Transactions on Biomedical Engineering* 57, 7 (2010), 1798–1806.
[40] Song Liu, Makoto Yamada, Nigel Collier, and Masashi Sugiyama. 2013. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks* 43 (2013), 72–83.
[41] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.
[42] Varun Mishra, Gunnar Pope, Sarah Lord, Stephanie Lewia, Byron Lowens, Kelly Caine, Sougata Sen, Ryan Halter, and David Kotz. 2020. Continuous detection of physiological stress with commodity hardware. *ACM transactions on computing for healthcare* 1, 2 (2020), 1–30.
[43] Prasanth Murali, Javier Hernandez, Daniel McDuff, Kael Rowan, Jina Suh, and Mary Czerwinski. 2021. Affectivespotlight: Facilitating the communication of affective responses from audience members during online presentations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
[44] Rab Nawaz, Kit Hwa Cheah, Humaira Nisar, and Vooi Voon Yap. 2020. Comparison of different feature extraction methods for EEG-based emotion recognition. *Biocybernetics and Biomedical Engineering* 40, 3 (2020), 910–926.
[45] Dan Nie, Xiao-Wei Wang, Li-Chen Shi, and Bao-Liang Lu. 2011. EEG-based emotion recognition during watching movies. In *2011 5th International IEEE/EMBS Conference on Neural Engineering*. IEEE, 667–670.
[46] Maria Augusta SN Nunes, Stefano A Cerri, and Nathalie Blanc. 2008. Towards user psychological profile. In *Proceedings of the VIII Brazilian symposium on human factors in computing systems*. 196–203.
[47] Cheul Young Park, Narae Cha, Soowon Kang, Auk Kim, Ahsan Habib Khandoker, Leontios Hadjileontiadis, Alice Oh, Yong Jeong, and Uichin Lee. 2020. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data* 7, 1 (2020), 1–16.
[48] Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2017. Searching for activation functions. *arXiv preprint arXiv:1710.05941* (2017).
[49] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.
[50] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.
[51] Luz Santamaria-Granados, Mario Munoz-Organero, Gustavo Ramirez-Gonzalez, Enas Abdulhay, and NJIA Arunkumar. 2018. Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS). *IEEE Access* 7 (2018), 57–67.
[52] Pritam Sarkar and Ali Etemad. 2020. Self-supervised ECG representation learning for emotion recognition. *IEEE Transactions on Affective Computing* (2020).
[53] Erich Schubert. 2023. Stop using the elbow criterion for k-means and how to choose the number of clusters instead. *ACM SIGKDD Explorations Newsletter* 25, 1 (2023), 36–42.
[54] Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52, 3/4 (1965), 591–611.
[55] Karan Sharma, Claudio Castellini, Freek Stulp, and Egon L Van den Broek. [n. d.]. Continuous, real-time emotion annotation: A novel joystick-based analysis framework. *IEEE Transactions on Affective Computing* 11, 1 ([n. d.]).
[56] Karan Sharma, Claudio Castellini, Egon L van den Broek, Alin Albu Schaeffer, and Friedhelm Schwenker. 2019. A dataset of continuous affect annotations and physiological signals for emotion analysis. *Scientific data* 6, 1 (2019).
[57] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. 2018. A review of emotion recognition using physiological signals. *Sensors* 18, 7 (2018), 2074.
[58] Jainendra Shukla, Miguel Barreda-Angeles, Joan Oliver, Gora Chand Nandi, and Domenec Puig. 2019. Feature extraction and selection for emotion recognition from electrodermal activity. *IEEE Transactions on Affective Computing* 12, 4 (2019), 857–869.
[59] Abhijeet Swain, Vaibhav Ganatra, Snehanshu Saha, Archana Mathur, and Rekha Phadke. 2022. P-LSTM: A Novel LSTM Architecture for Glucose Level Prediction Problem. In *International Conference on Neural Information Processing*. Springer, 369–380.
[60] Luma Tabbaa, Ryan Searle, Saber Mirzaee Bafti, Md Moinul Hossain, Jittrapol Intarasisrisawat, Maxine Glancy, and Chee Siang Ang. 2021. Vreed: Virtual reality emotion recognition dataset using eye tracking & physiological measures. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5, 4 (2021), 1–20.
[61] Hao Tang, Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. 2017. Multimodal emotion recognition using deep neural networks. In *International Conference on Neural Information Processing*. Springer, 811–819.
[62] Philippe Verduyn, Ellen Delvaux, Hermina Van Coillie, Francis Tuerlinckx, and Iven Van Mechelen. 2009. Predicting the duration of emotional experience: two experience sampling studies. *Emotion* 9, 1 (2009), 83.

[63] Philippe Verduyn and Saskia Lavrijsen. 2015. Which emotions last longest and why: The role of event importance and rumination. *Motivation and Emotion* 39, 1 (2015), 119–127.
[64] Julia Wache, Ramanathan Subramanian, Mojtaba Khomami Abadi, Radu-Laurentiu Vieriu, Nicu Sebe, and Stefan Winkler. 2015. Implicit user-centric personality recognition based on physiological responses to emotional videos. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 239–246.
[65] Zhongmin Wang, Xiaoxiao Zhou, Wenlang Wang, and Chen Liang. 2020. Emotion recognition using multimodal deep learning in multiple psychophysiological signals and video. *International Journal of Machine Learning and Cybernetics* 11, 4 (2020), 923–934.
[66] Axel Wismüller, Adora M Dsouza, M Ali Vosoughi, and Anas Abidin. 2021. Large-scale nonlinear Granger causality for inferring directed dependence from short multivariate time-series data. *Scientific reports* 11, 1 (2021), 7817.
[67] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2, 1-3 (1987), 37–52.
[68] Chi-Keng Wu, Pau-Choo Chung, and Chi-Jen Wang. 2012. Representative segment-based emotion analysis and classification with automatic respiration signal segmentation. *IEEE Transactions on Affective Computing* 3, 4 (2012), 482–495.
[69] Shanxiao Yang and Guangying Yang. 2011. Emotion Recognition of EMG Based on Improved LM BP Neural Network and SVM. *J. Softw.* 6, 8 (2011), 1529–1536.
[70] Tianyi Zhang, Abdallah El Ali, Chen Wang, Alan Hanjalic, and Pablo Cesar. 2020. RCEA: Real-time, Continuous Emotion Annotation for Collecting Precise Mobile Video Ground Truth Labels. In *ACM CHI*. 1–15.

A Appendix

A.1 Sensor Signal Quality Analysis

In this section, we investigate the sensor (HR, and GSR) signal quality. To perform this, we compared the GSR V1.2 sensor (Seeed Studio Grove) and the pulse rate sensor (HW-827, World Famous Electronics LLC) with the Empatica Embrace Plus. EDA and pulse rate data were collected from the Empatica Embrace Plus while GSR and HR values were recorded from our setup as participants watched stimuli videos. The cosine similarity between the pulse rate from the Empatica Embrace Plus and the HR sensor was 0.99, and between the EDA from the Empatica Embrace Plus and the GSR sensor was also 0.99. These high similarity values indicate that the data from these sensors are comparable to the accurate readings of the Empatica wristband, and we preferred using these sensors due to challenges with the real-time processing of Empatica data.

A.2 Algorithm for Labelling Opportune Moments

We outline the steps of labeling the opportune moments in Algorithm 1. At first, we pass the segmented physiological signals as input to the algorithm and compute the change point score (`line 1-5`). Next, we calculate the mean and SD (`line 6-8`) of the change point score. Then we find the outliers at the higher end, which are considered as opportune moments from the calculated change point score (`line 9-14`). Later, we identify non-outliers and perform k-means clustering (k=2)(`line 17-22`) on it. Next, pick the cluster with a higher centroid value, and identify only those points within the cluster having a value higher than the centroid (`line 23-33`). These identified points are also marked as opportune moments (as they indicate a significant change in the physiological responses).

A.3 Video-wise Emotion Annotation Quality Comparison

In this section, we show that there is no significant difference in the video-wise valence (and arousal) scores between ground truth continuous data and the probed values sampled using the PResUP framework. In this case also, we repeat the experiments twice (once for valence, once for arousal).

In specific, we performed the following steps for every video. First, we computed the average valence value from the continuous annotations for every video. Then, we found the average value of valence from only the probed segments of the corresponding video. We checked the distribution of these values to verify if they follow normal distribution. We performed the Shapiro-Wilk test [54], which revealed that the distribution is not normal.

Algorithm 1: Algorithm for labeling opportune moment in the dataset

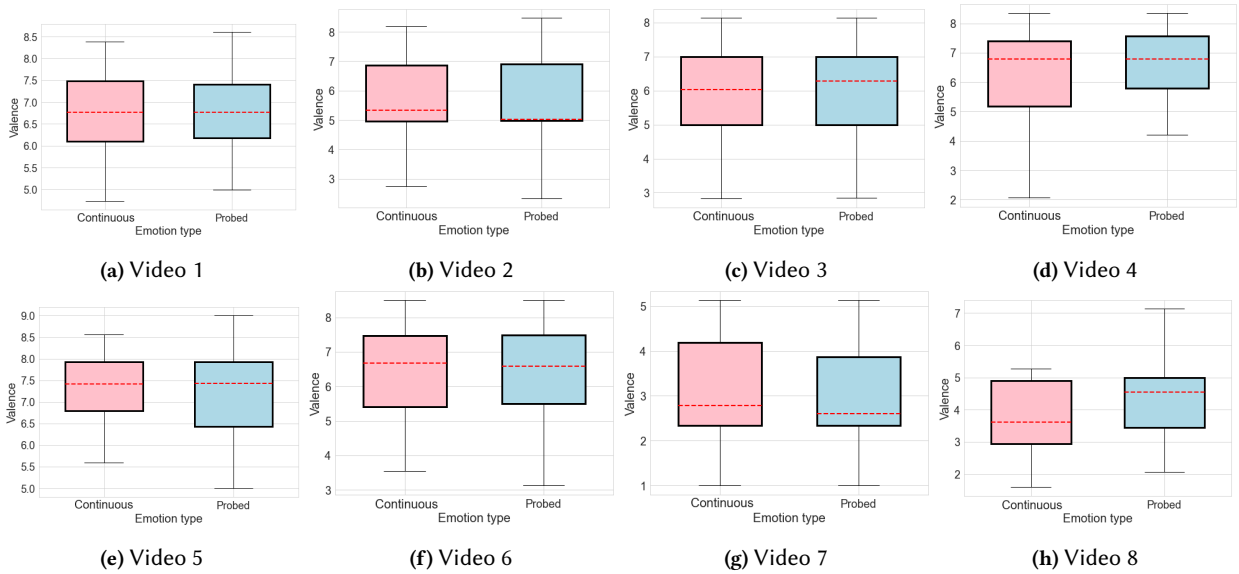
Input: $\mathbb{PS} = \{[PS]_1, .. [PS]_{N-1}\}$, segmented physiological signals.
Output: P: Opportune moments
/* Calculate change point scores */
for $i \leftarrow 1$ ***to*** $(n-1)$ **do**
  $[S]^i \leftarrow$ RuLSIF(PS[i-1], PS[i])
  $[S] \leftarrow [S] \cup [S]^i$
/* Calculate the mean, SD of change point scores */
$\mu_s \leftarrow$ mean(S)
$\sigma_s \leftarrow$ standard_deviation(S)
/* Calculate Outliers at higher end */
for $i \leftarrow 1$ ***to*** $(n-1)$ **do**
  if $[S]^i \geq \mu_s + (3 * \sigma_s)$ **then**
    $O_s \leftarrow [S]^i$
/* outliers are considered as opportune moment */
$P \leftarrow O_s$
/* Identifying Non-Outliers */
$NO_s \leftarrow [S] - O_s$
/* Set the value of k for k-means clustering */
$k \leftarrow 2$
/* Perform k-means clustering */
$[cluster, c\_ids] \leftarrow$ k_means_clustering(NO_s, k)
/* Finding the cluster with the higher centroid */
$[cd1, cd2] \leftarrow$ Compute_centroid($cluster, c\_ids$)
$[cluster\_of\_interest, cd] \leftarrow$ Max($cd1, cd2$)
/* Identifying opportune moment from Non-Outliers */
for $i \leftarrow 1$ ***to*** $(n-1)$ **do**
  if $[C\_ids]^i = cluster\_of\_interest$ **then**
    if $[NO_s]^i \geq cd$ **then**
      $P \leftarrow [NO_s]^i$
return P

As a result, we performed the Mann-Whitney U test [41] to compare if there is a significant difference in the valence scores between ground truth annotations and sampled annotations. We did not observe a significant difference for each of the videos in the valence scores present in continuous annotations and collected using the PResUP framework (see Fig. 14). The same steps were performed for video-wise arousal score comparison and in this case also, we did not observe a significant difference for any video (see Fig. 15). These results demonstrate that there is no significant difference in video-wise valence (and arousal) scores of continuous annotation data and probed annotations.

A.4 Generalization of PResUP: CASE Dataset

We evaluated the generalizability of PResUP on the publicly available CASE [56] dataset. The dataset captures the physiological responses from six sensor streams from 30 participants as they watch eight different videos and perform the continuous annotation using a joystick. The continuous annotations were collected on a 5-point valence-arousal scale based on the Circumplex model [50] of the emotion. We implemented the pre-processing steps outlined in Section 3.3, which yielded a total of 7290 segments (11.6% opportune, 88.4% inopportune), thus, on average, 243 segments per user. We applied PResUP on this processed dataset to detect the opportune moments.

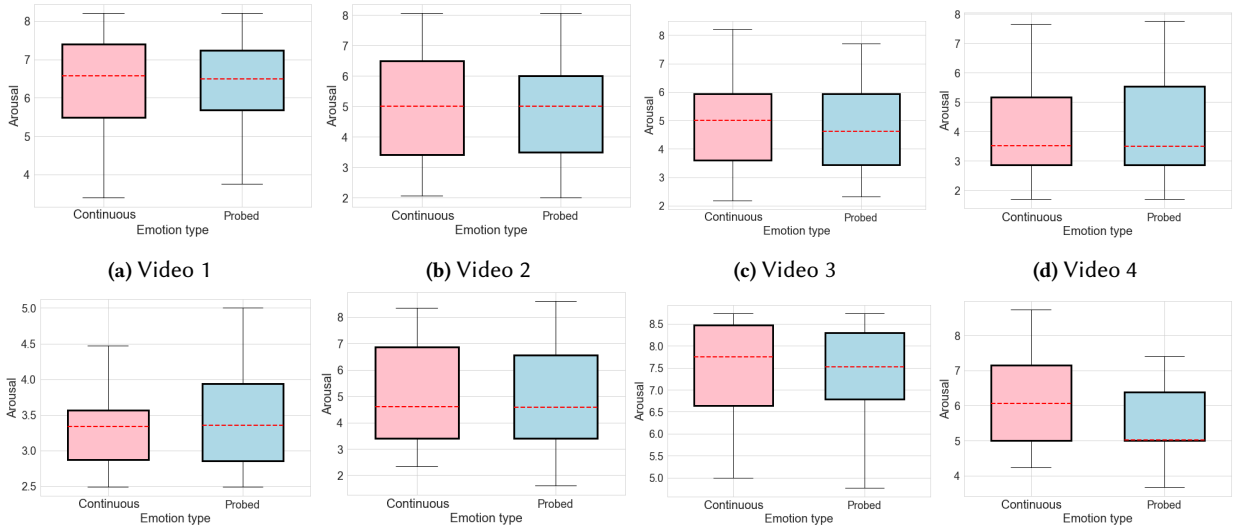
The performance comparison of PResUP and the baselines on the CASE dataset is presented in Table 7. We note that PResUP outperforms all the baselines in terms of probing rate (5.80 (per user per video)), therefore offers an average reduction of 26.19% in probing rate compared to the baselines. It also has the highest LR+ (7.01),



(a) Video 1 **(b)** Video 2 **(c)** Video 3 **(d)** Video 4

(e) Video 5 **(f)** Video 6 **(g)** Video 7 **(h)** Video 8

Fig. 14. Comparison of video-wise valence scores between the ground truth and probed moments' annotation reveals that there is no significant difference between the continuous and probed values. Mann-Whitney U test shows valence scores do not vary significantly ($p < 0.05$) between the continuous and probed annotations for the videos.



(a) Video 1 **(b)** Video 2 **(c)** Video 3 **(d)** Video 4

(e) Video 5 **(f)** Video 6 **(g)** Video 7 **(h)** Video 8

Fig. 15. Comparison of video-wise arousal scores between the ground truth and probed moments' annotation reveals that there is no significant difference between the continuous and probed values. Mann-Whitney U test shows arousal scores do not vary significantly ($p < 0.05$) between the continuous and probed annotations for the videos.

the highest mean TPR (86.07%), and the least mean FPR (12.27%). All these findings highlight the utility of PResUP in reducing the continuous emotion annotation effort without compromising the annotation quality.

	Probing rate↓	TPR (%) ↑	FPR (%) ↓	LR+ ↑
TBS	30.38 (0.00)	100.00 (0.00)	100.00 (0.00)	1.00
RPS	15.27 (0.01)	52.35 (0.01)	49.98 (0.01)	1.05
PPS	20.38 (1.33)	69.21 (32.34)	16.16 (18.72)	4.28
FBS	6.50 (3.02)	70.49 (22.49)	15.02 (8.31)	4.69
APS	6.10 (4.04)	71.18 (20.86)	14.35 (15.98)	4.96
GBPS	6.07 (4.07)	82.12 (16.18)	13.68 (13.19)	6.00
GPS	6.37 (4.10)	81.30 (21.24)	14.93 (14.15)	5.45
RNNPS	6.01 (8.72)	82.66 (11.52)	12.88(5.24)	6.41
GRUPS	6.03 (8.98)	76.14 (10.75)	13.68 (5.37)	5.56
CNNPS	6.52 (6.76)	82.55(10.39)	15.47(6.96)	5.33
PResUP	**5.80 (2.89)**	**86.07 (11.64)**	**12.27 (9.67)**	**7.01**

Table 6. Performance comparison of PResUP and baselines on the CASE [56] dataset. The values indicate the user-wise average and std.dev (inside parenthesis) for a metric. PResUP outperforms all baselines in terms of probing rate (least probing rate), LR+ (highest LR+), TPR, and FPR. Although **TBS** has the highest TPR, it also has the worst FPR (100%). ↑, ↓ indicate higher and lower value preferred respectively.

A.5 Generalization of PResUP: K-Emocon Dataset

	Probing rate↓	TPR (%) ↑	FPR (%) ↓	LR+ ↑
TBS	237.28 (0.00)	100 (0.00)	100 (0.00)	1.00
RPS	118.48 (0.01)	50.88 (0.01)	49.76 (0.01)	1.02
PPS	74.46 (14.87)	81.26 (15.41)	13.01 (1.02)	6.25
FBS	32.32 (16.61)	80.22 (16.19)	4.04 (5.86)	19.86
APS	33.82 (16.41)	78.28 (17.51)	3.72 (5.74)	21.04
GBPS	33.78 (17.22)	80.01 (17.16)	3.74 (5.98)	21.39
GPS	33.89 (21.96)	75.65 (36.17)	5.16 (6.58)	14.66
RNNPS	34.35 (23.93)	68.39 (44.94)	6.56 (6.92)	10.42
GRUPS	40.96 (18.35)	50.51 (43.57)	11.60 (2.94)	4.35
CNNPS	44.5 (16.38)	73.53 (27.01)	11.41 (5.95)	6.44
PResUP	**31.36 (11.23)**	**82.26 (13.38)**	**3.41 (4.41)**	**24.12**

Table 7. Performance comparison of PResUP and baselines on K-emocon [47] dataset. The values indicate the user-wise average and std.dev (inside parenthesis) for a metric. PResUP outperforms all baselines in terms of probing rate (least probing rate), LR+ (highest LR+), TPR, and FPR. Although **TBS** has the highest TPR, it also has the worst FPR (100%). ↑, ↓ indicate higher and lower value preferred respectively.

We evaluated the generalizability of PResUP on the publicly available K-emocon [47] dataset. This is an audio-visual dataset that records physiological responses from three wearable devices during 16 paired debates (a total of 32 participants) on a social issue. The continuous annotations were collected on a 5-point valence-arousal scale as per the Circumplex model [50] of the emotion. We implemented the pre-processing steps outlined in Section 3.3, which yielded a total of 6644 segments (13.6% opportune, 86.3% inopportune), thus on average 238 segments per user. We applied PResUP on this processed dataset to detect the opportune moments.

The performance comparison of PResUP and the baselines on the K-emocon dataset is presented in Table 7. We note that PResUP outperforms all the baselines in terms of probing rate (31.36 (per user)), therefore offers an average reduction of 30.47% in probing rate compared to the baselines. It also has the highest LR+ (24.12), the highest mean TPR (82.26%), and the least mean FPR (3.41%). All these findings highlight the utility of PResUP in reducing the continuous emotion annotation effort without compromising the annotation quality.