Change-Point Detection in Emotion Recognition with Physiological Data

Dalkiran, Harun
Chair for Communication Technology
University of Kassel
Kassel, Germany
harundalkiran@uni-kassel.de

Noori, Ahmad Lowejatan
Chair for Communication Technology
University of Kassel
Kassel, Germany
l.noori@uni-kassel.de

***Abstract*—Emotion recognition from physiological signals has emerged as a key component in affective computing and human-robot interaction. This paper investigates two unsupervised change-point detection algorithms — one based on a kernel method (*iCID*) and the other on random forests (*ChangeForest*) — to identify emotional shifts in physiological data. Using the AFFECT-HRI dataset, we analyze Blood Volume Pulse (BVP) and Galvanic Skin Response (GSR) signals from 22 participants during interactions with a robot. Extensive preprocessing, including statistical and physiological feature extraction, was applied. The models are evaluated using precision, recall, F1-score, and mean absolute error relative to annotated emotion transitions. Results show that ChangeForest achieves higher precision and recall than iCID, particularly on BVP signals, whereas iCID demonstrates more balanced performance across both signal types. Although both approaches yield modest F1 scores, they align with expectations in the field and demonstrate the feasibility of detecting affective state changes through physiological signals. Beyond the comparative evaluation, this work highlights the broader challenges of applying unsupervised methods in noisy, real-world physiological data.**

I. Introduction

Physiological signals such as blood volume pulse (BVP) and galvanic skin response (GSR) are widely used in affective computing, human robot interaction (HRI), and stress monitoring as non-invasive indicators of emotional states [1], [2]. GSR is considered a reliable marker of arousal, while BVP, through heart rate and variability, reflects both sympathetic and parasympathetic activity [3]. Since these reactions are mostly involuntary, they offer an objective view of changes in arousal and emotion that people cannot easily control [4]. This has made BVP and GSR valuable in applications ranging from wearable stress detection to adaptive HRI systems.

A central challenge, however, is to detect when emotional shifts actually occur in these signals. Such change points often indicate transitions in affective or cognitive state, but most existing methods rely on supervised learning with fixed windows and dense labels. This is problematic: annotations are expensive and subjective, people differ in their baseline physiology, and fixed windows risk missing rapid transitions or blending multiple states [4]. As a result, supervised approaches are often difficult to scale beyond controlled laboratory settings.

Unsupervised change point detection (CPD) provides a promising alternative. Instead of labels, these methods look for moments where the statistical properties of the signal change [5], [6]. By exploiting the data structure itself, CPD can monitor affective dynamics more continuously and in real time.

In this paper, we compare two state-of-the-art unsupervised CPD algorithms. The first, *iCID*, uses isolation kernels to detect shifts between adjacent signal segments [5]. The second, *ChangeForest*, applies random forests to partition multivariate time series [6]. Using the AFFECT-HRI dataset [2], which contains BVP and GSR recordings, we evaluate how well these methods align with self-reported emotional changes. We report precision, recall, F1 score, and mean absolute error, and discuss the strengths of kernel- versus tree-based approaches for real-world emotion recognition.

II. Related Work

Time series physiological signals correlate with human affective states, and change points indicate shifts in emotional state. Such signals can be acquired via wearable sensors, with ground-truth labels often collected through participant self-reports during or after experiments.

Adithya et al. [7] developed OCEAN to reduce annotation effort and cognitive workload. They segmented physiological signals into windows, used the RuLSIF density-ratio algorithm to compute change-point scores, and applied k-means clustering to select relevant self-report moments, showing high similarity between probed and sampled valence and arousal annotations.

Avramidis et al. [8] used driving experiment data, including Breath Rate and EDA, as stress indicators. They combined Greedy Gaussian Segmentation (GGS) with k-means clustering to discard similar change points, achieving above 70

Arabian et al. [9] detected emotional events offline from EDA and ECG during emotion-inducing images. After preprocessing with piecewise-cubic interpolation and setting negative values to zero, they identified stimulus onset via peaks in the second derivative of EDA, achieving a root mean square error of 0.99 seconds.

Hamidi et al. [10] applied Bayesian change-point detection and dynamic time-warping to stress-inducing tasks. Change

points segmented physiological data, and segment similarity within the same condition validated alignment with stress transitions, though finer emotion classes were not analyzed.

Despite these advances, challenges remain: emotional labels are subjective, and self-reports can be affected by cognitive workload and timing. Physiological responses vary across individuals, may be delayed or subtle, and wearable sensors introduce noise and artifacts, increasing false positives. Methods often rely on fixed window sizes or strong signal assumptions, where small windows weaken statistical power and large windows may mix distributions [11].

III. Change-Point Detection Algorithms

In this section, we present two unsupervised learning algorithms for change-point detection. First, we formally define what change points are. Then, we introduce the algorithms.

A. Change Points

A *change point* marks a time index at which the underlying distribution generating a data sequence changes, often abruptly.

Let $x_1, \ldots, x_N$ be a sequence of d-dimensional observations, where $x_i \in \mathbb{R}^d$. Suppose the sequence is partitioned into K segments at unknown indices $t_1, \ldots, t_{K-1}$, with $t_0 = 1$ and $t_k = N$. Let each segment $[x_{t_i}, x_{t_{i+1}}]$ consist of i.i.d. samples drawn from a distribution P_i with $P_i \neq P_{i+1}$ for all $i \in \{1, \ldots, K-1\}$. Change-point detection consists of estimating the unknown indices $t_1, \ldots, t_{k-1}$ at which the data distribution changes.

B. Change-Point Detection with the Isolation Distributional Kernel

First, we used the change-point detection algorithm based on an Isolation Distributional Kernel, called *iCID* and originally proposed by Cao et al. [12] for change-interval detection. Intuitively, a distributional kernel measures similarity between distributions via a specific feature map. A point x_t is classified as a change-point if the corresponding dissimilarity between its adjacent windows exceeds some threshold value.

Let ψ be the number of center points uniformly sampled from the time-series $D = \{x_1, \ldots, x_N\}$. Each point in D is assigned to its nearest center, forming a partition of D into ψ subsets. The feature map Φ of an isolation distributional kernel is a map $\Phi : \mathbb{R} \to \{0,1\}^{t \cdot \psi}$, where $\Phi(x)$ is a concatenation of t one-hot encoded vectors, each indicating x's nearest center point in each of the t random partitions. The similarity of two distributions P_X and P_Y can be measured by the normalised Isolation Distributional Kernel (IDK)

$$\mathcal{K}_I(P_X, P_Y|D) = \frac{\frac{1}{t}\langle \hat{\Phi}(P_X|D), \hat{\Phi}(P_Y|D)\rangle}{\|\hat{\Phi}(P_X|D)\| \|\hat{\Phi}(P_Y|D)\|}, \tag{1}$$

where $\hat{\Phi}(P_X|D) = \frac{1}{t}\sum_{x \in X} \Phi(x|D)$ is the kernel mean embedding and $\Phi(x|D)$ the feature map from before. The dissimilarity score $\mathfrak{S}(X, Y)$ of two intervals X and Y is then naturally $\mathfrak{S}(X, Y) = 1 - \mathcal{K}_I(P_X, P_Y)$.

Assume a fixed window size w and let the past interval $X_{<t} := \{x_{t-w}, \ldots, x_{t-1}\}$ and the current interval $X_{>t} := \{x_{t+1}, \ldots, x_{t+w}\}$ at each point x_t. The change-point is the x_t which has a large score more than a threshold τ, i.e. $\mathfrak{S}(X_{<t}, X_{>t}) \geq \tau$. To reduce the false positives rate, only local maxima of points above the threshold are selected, also known as *non-maximum suppression* [13].

The key parameter ψ is selected based on the stability of the computed scores. Let $C_\psi := \{\mathfrak{S}(X_{<i}, X_{>i}) \mid i \in [w+1, N-w], \psi\}$ be the set of all scores dependent on ψ in the feature map of the IDK. The optimal ψ leads to a minimum of instability, hence

$$\psi^* = \arg\min_{\psi} \bar{E}(C_\psi), \tag{2}$$

where $\bar{E}$ is a measure of instability, e.g. approximated entropy estimation or variance. The threshold τ can be identified as

$$\tau(C_\psi) = \mu + \alpha \cdot \sigma, \tag{3}$$

where μ and σ are mean and standard deviation of C_{ψ^*}, respectively, and α a power factor parameter.

C. Change-Point Detection with Random Forest

We additionally applied the *ChangeForest* algorithm, as proposed by Londschien et al. [14]. ChangeForest is a decision tree-based method for multiple change-point detection in multivariate time series. The main idea is to recursively partition the time series into segments where the statistical properties of the data remain as consistent as possible. At each node of the recursive partitioning, the algorithm evaluates all possible split points within the current interval and selects the split with the highest statistical gain.

Formally, for an interval (u, v) and a candidate split point s with $u < s < v$, the gain is defined as:

$$\begin{aligned} G_{(u,v)}(s) = &\sum_{i=u+1}^{s} \log \frac{p_{\hat{\theta}_{(u,s)}}(x_i)}{p_{\hat{\theta}_{(u,v)}}(x_i)} \\ &+ \sum_{i=s+1}^{v} \log \frac{p_{\hat{\theta}_{(s,v)}}(x_i)}{p_{\hat{\theta}_{(u,v)}}(x_i)} \end{aligned} \tag{4}$$

Here, $p_{\hat{\theta}_{(u,s)}}$ and $p_{\hat{\theta}_{(s,v)}}$ denote the maximum-likelihood probability density functions fitted to the subsequences (u, s) and (s, v), respectively, under the assumption of multivariate normality. The parameter vectors $\hat{\theta}$ consist of the estimated mean vector and covariance matrix for the data in each segment. Likewise, $p_{\hat{\theta}_{(u,v)}}$ is the density estimated from the entire interval before the split. The intuition is that if s is a true change-point, then fitting two separate distributions (before and after s) explains the data significantly better than fitting a single distribution to the entire interval. The gain function thus quantifies the improvement in model fit by introducing a split at s.

To avoid overfitting and unnecessary segmentation, ChangeForest complements the gain criterion with statistical hypothesis testing. For each candidate split s, the null hypothesis and alternative hypothesis are defined as:

$$H_0 : \text{no change at } s \quad \text{vs.} \quad H_1 : \text{change at } s \tag{5}$$

The corresponding test yields a p-value $p(s)$. Only if this p-value satisfies

$$p(s) < \alpha, \tag{6}$$

where α is a predefined significance level (e.g., 0.05), is the change-point accepted. This statistical safeguard ensures that the method controls the false positive rate and only introduces splits when there is strong evidence of a distributional change.

The recursive application of this splitting and testing procedure results in a binary segmentation tree, where each node corresponds to a homogeneous segment and each internal node represents a detected change-point. Because the method is data-driven and does not require prior knowledge of the number of change-points, it can flexibly adapt its complexity to the underlying structure of the time series. Additionally, by relying on random forests to approximate likelihood ratios in high-dimensional feature spaces, ChangeForest can handle complex, multivariate, and non-stationary data. This makes it a robust and versatile tool for real-world time-series applications, including physiological signal analysis.

IV. Methodology

In this section, we will discuss data analysis and preprocessing steps.

A. Data Analysis

We used the AFFECT-HRI dataset [2], which includes physiological signals from 146 participants recorded via an E4 wristband during human-robot interaction. Perceived emotions at certain moments were derived from post-experiment questionnaires. For this study, we focused on neutral scenarios with the Tiago++ robot, limiting the data to 22 participants and using only Blood Volume Pulse (BVP) and Galvanic Skin Response (GSR) signals, sampled at 64 Hz and 4 Hz, respectively.

B. Data Preprocessing

Physiological signals often contain noise, outliers, or missing values. We preprocessed GSR and BVP with *NeuroKit2* [15]; missing values were forward-filled (last valid reading). Because BVP is analogous to PPG [16], we applied the standard PPG cleaning pipeline: a 3rd-order Butterworth bandpass (0.5–8 Hz) to suppress high-frequency noise [17] and clipping to the 1st–99th percentiles to reduce noise sensitivity [18]. GSR's low sampling rate received no frequency filtering. From GSR, we extracted tonic and phasic components, and from BVP we derived heart rate, all via *NeuroKit2*. For both signals we computed a 1-second sliding-window mean and standard deviation, and then standardized all features.

When ground-truth files lacked emotion annotations at labeled HRI moments, we populated them from `questionnaire.csv` (valence–arousal ratings mapped to the valence–arousal plane of [2]). Consecutive identical labels kept only the first emotion entry, while non-consecutive repetitions (e.g., robot repetitions) were preserved. Finally, BVP and GSR were order-merged with emotion labels using NTPTime timestamps and trimmed from HRI onset to the robot's final sentence.

V. Evaluation

In this section, we compare both models on the physiological signals of 22 participants from the AFFECT-HRI dataset. The algorithms are applied to each BVP and GSR signal. By analyzing both signals, we aim to capture a broader range of emotional changes. The ground-truth change-points correspond to time steps where an emotion was perceived. These are determined by five questions in the post-questionnaire linked to specific sentences spoken by the robot. Therefore, also considering repetitions and missing sentences, there are at most 5–6 change-points per participant. In each question, the participant was required to enter a valence and arousal value between 1 and 5. Through a valence-arousal-plane, these values were categorized to an emotion: HNVHA as stressed, HPVHA as happy, HNVLA as depressed, and HPVLA as relaxed [2]. The entire interaction lasts approximately 4 minutes per participant.

TABLE I
Parameter settings for iCID and ChangeForest

Algorithm	**Parameters**
iCID	$t = 200$; $\psi \in \{2, 4, 8, 16, 32, 64\}$ selected via internal grid search; $\bar{E}$ = variance; GSR: $w = 50$ (12.5 s), $\alpha = 1$; BVP: $w = 200$ (3.125 s), $\alpha = 2$;
ChangeForest	Minimal relative segment length = 0.05; Minimal gain to split = 100; Gain threshold = 100; Number of random forest estimators = 100;

TABLE II
Evaluation Metrics for a 10-second Ground-Truth Window

Ground Truth Window (s)	10			
Metric	**iCID**		**ChangeForest**	
	GSR	**BVP**	**GSR**	**BVP**
Micro Precision	0.301	0.290	0.473	0.362
Micro Recall	**0.715**	**0.788**	**0.715**	**0.967**
Micro Avg. F1	0.424	0.424	0.570	0.527
Macro Precision	0.285	0.246	0.405	0.311
Macro Recall	0.659	0.705	0.671	0.953
Macro Avg. F1	0.370	0.358	0.501	0.466

Table II summarizes evaluation metrics using a 10-second ground-truth window with the parameter settings given by I. Following Yu et al. [11], predicted change-points are counted as true positives if they occur within this tolerance of a ground-truth annotation. Micro metrics are computed globally across

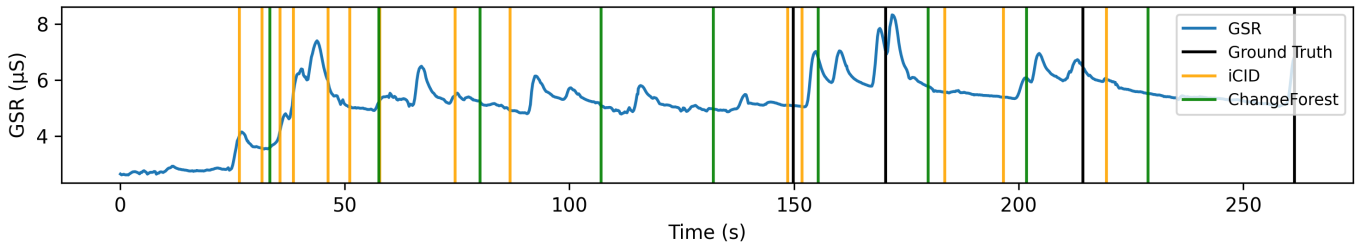

Fig. 1. Visual comparison of detected change-points by iCID and ChangeForest on the GSR signal of one participant.

participants, aggregating true positives, false positives, and false negatives. In contrast, macro metrics average performance across all participants. The 10-second window was chosen to account for the typical 3–15 second latency of BVP and GSR responses [19].

Overall, *ChangeForest* achieves higher precision and recall, resulting in superior F1-scores. The F1-scores are particularly high in the GSR data for *ChangeForest* (see Table II). The metrics for GSR and BVP are more balanced for *iCID* than for *ChangeForest*. Both algorithms perfom the highest in the micro recall. It is evident that *iCID* tends to detect many change-points in clusters, whereas *ChangeForest* identifies them more sparsely. This behavior increases the likelihood that *ChangeForest*'s detections fall within the 10-second ground-truth window, resulting in a high true positive rate. A similar pattern can be observed for a sample participant in Figure 1.

TABLE III
MEAN ABSOLUTE ERROR ACROSS ALL PARTICIPANTS

Metric	**iCID**		**ChangeForest**	
	GSR	**BVP**	**GSR**	**BVP**
MAE (s)	17.34	**16.71**	**16.89**	17.59

Table III reports the mean absolute error (MAE) of detected change-points. Both methods yield similar timing accuracy, with *ChangeForest* performing slightly better on GSR and *iCID* on BVP.

TABLE IV
EMOTION-WISE RECALL FOR CHANGE-POINT DETECTION

Emotion	**iCID**		**ChangeForest**	
	GSR	**BVP**	**GSR**	**BVP**
Stressed	0.68	0.74	0.42	0.89
Happy	0.79	0.86	0.71	**1.00**
Relaxed	**0.80**	**0.88**	**0.96**	0.96
Depressed	0.67	0.67	0.67	**1.00**
Neutral	0.50	0.62	0.62	**1.00**

Table IV shows emotion-wise recall for *iCID* and *ChangeForest*. *ChangeForest* achieves near-perfect recall on BVP across all emotions, while its GSR performance is lower for high-arousal states (Stressed: 0.42; Happy: 0.71). *iCID* shows more balanced recall across GSR and BVP, performing well on Relaxed and Happy states but lower on Neutral and Stressed. Overall, *ChangeForest* excels on BVP, whereas *iCID* maintains more consistent performance on GSR.

TABLE V
COMPUTATION TIME FOR EACH METHOD ON PREPROCESSED BVP DATA OF ONE PARTICIPANT (300,000 × 5) USING AN AMD RYZEN 5 5500U (16 GB RAM)

Method	**Runtime (s)**
ChangeForest	**1.68**
iCID with Approx. Entropy	5106.45
iCID with Variance	40.07

ChangeForest is efficient and scales linearly, even with high-dimensional data [14], using gain-based binary segmentation. In contrast, *iCID* evaluates scores across nearly all points and suffers from high computational cost (see Table V).

VI. CONCLUSION

This study evaluated the effectiveness of a kernel-based and a random forest-based change-point detection algorithm for identifying emotional shifts using physiological signals, specifically BVP and GSR. Preprocessing involved extracting statistical features (e.g., mean, standard deviation) with a sliding window and physiological features (e.g., heart rate from BVP, tonic/phasic components from GSR). The models were benchmarked using standard evaluation metrics, with emphasis on recall to assess alignment with ground-truth emotional states. Runtime performance was also measured on long BVP signals to examine real-time feasibility.

The results indicate that the Random Forest-based approach is better suited for recognizing emotional states in physiological data. It is important to note that results in Table II, III and IV are affected by uncertainties in ground-truth labeling. In our case, self-reported post-questionnaires may not fully capture participants' actual states due to memory decay, cognitive load, or misalignment between perceived emotions and marked HRI events. While achieved F1 scores ranged between 0.35 and 0.57, these values are consistent with recent literature for similar tasks [11], reflecting the inherent complexity of change-point detection. We also acknowledge limited statistical power due to sample size. Future work should address this by including more participants and exploring additional data modalities to improve generalizability.

REFERENCES

[1] Mikel Val-Calvo, José R. Álvarez-Sánchez, José M. Ferrández-Vicente, and Eduardo Fernández. Affective robot story-telling human-robot interaction: Exploratory real-time emotion estimation analysis using facial expressions and physiological signals. *IEEE Access*, 8:134051–134066, 2020.

[2] Judith S. Heinisch, Jérôme Kirchhoff, Philip Busch, Janine Wendt, Oskar von Stryk, and Klaus David. Physiological data for affective computing in hri with anthropomorphic service robots: the affect-hri data set. *Scientific Data*, 11(1):333, 2024.

[3] Chin-An Wang, Talia Baird, Jeff Huang, Jonathan D. Coutinho, Donald C. Brien, and Douglas P. Munoz. Arousal effects on pupil size, heart rate, and skin conductance in an emotional face task. *Frontiers in Neurology*, 9:1029, 2018.

[4] Mengting Zhang and Yixuan Cui. Self supervised learning based emotion recognition using physiological signals. *Frontiers in Human Neuroscience*, 18:1334721, 2024.

[5] Yang Cao, Ye Zhu, Kai Ming Ting, Flora D. Salim, Hong Xian Li, Luxing Yang, and Gang Li. Detecting change intervals with isolation distributional kernel. *Journal of Artificial Intelligence Research*, 79:273–306, 2024.

[6] Malte Londschien, Peter Bühlmann, and Solt Kovács. Random forests for change point detection. *Journal of Machine Learning Research*, 24:1–45, 2023.

[7] Akhilesh Adithya, Snigdha Tiwari, Sougata Sen, Sandip Chakraborty, and Surjya Ghosh. 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)*, pages 9–12, 2022.

[8] Kleanthis Avramidis, Tiantian Feng, Digbalay Bose, and Shrikanth Narayanan. Multimodal estimation of change points of physiological arousal during driving. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2023 - Workshops, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE, 2023.

[9] Herag Arabian, Ramona Schmid, Verena Wagner-Hartl, and Knut Moeller. Analysis of eda and heart rate signals for emotional stimuli responses. *Current Directions in Biomedical Engineering*, 9(1):150–153, 2023.

[10] Hossein Hamidi Shishavan, Ethan Gossett, Jinbo Bi, Robert Henning, Martin Cherniack, and Insoo Kim. Unsupervised bayesian change point detection model to track acute stress responses. *Biomedical Signal Processing and Control*, 95:106415, 2024.

[11] Jennifer Yu, Tina Behrouzi, Kopal Garg, Anna Goldenberg, and Sana Tonekaboni. Dynamic interpretable change point detection for physiological data analysis. In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh, editors, *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 636–649. PMLR, 10 Dec 2023.

[12] Yang Cao, Ye Zhu, Kai Ming Ting, Flora D. Salim, Hong Xian Li, Luxing Yang, and Gang Li. Detecting change intervals with isolation distributional kernel. *J. Artif. Int. Res.*, 79, April 2024.

[13] Z Ebrahimzadeh and S Kleinberg. Multi-scale change point detection in multivariate time series. In *NIPS Time Series Workshop*, 2017.

[14] Malte Londschien, Peter Bühlmann, and Solt Kovács. Random forests for change point detection, 2023.

[15] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, feb 2021.

[16] John Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 28(3):R1, feb 2007.

[17] Alina Nechyporenko, Marcus Frohme, Yaroslav Strelchuk, Vladyslav Omelchenko, Vitaliy Gargin, Liudmyla Ishchenko, and Victoriia Alekseeva. Galvanic skin response and photoplethysmography for stress recognition using machine learning and wearable sensors. *Applied Sciences*, 14(24), 2024.

[18] Patrícia J. Bota, Chen Wang, Ana L. N. Fred, and Hugo Plácido Da Silva. A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals. *IEEE Access*, 7:140990–141020, 2019.

[19] Kalliopi Kyriakou, Bernd Resch, Günther Sagl, Andreas Petutschnig, Christian Werner, David Niederseer, Michael Liedlgruber, Frank H. Wilhelm, Tess Osborne, and Jessica Pykett. Detecting moments of stress from measurements of wearable physiological sensors. *Sensors*, 19(17), 2019.