Personal and Contextual Knowledge Driven Robust Multimodal Affect Recognition using Smartwatches

Sirat Samyoun* and John Stankovic

Abstract— **Accurate and robust affect recognition in the wild is challenging using smartwatches due to scarcity of labeled sensor data. Although smartwatches can easily collect additional information such as, personal and contextual attributes related to affective events, the existing models fail to extract useful representations from such information and thus suffer from performance degradation under various settings. To tackle this problem, we present a novel multimodal machine learning framework that utilizes representation from the personal and contextual attributes as well as from limited sensor data. A real-life user study with 19 participants, followed by extensive evaluation shows that our solution outperforms the existing works across various affective tasks and improves the generalizability of the affective models.**

I. Introduction

Several mental health conditions including anxiety, depression, chronic stress have risen alarmingly worldwide in recent years [1]. The complex nature of these conditions leads to higher suicide and mortality rates, and significant economic burden. Recent research in affective computing, empowered by the smartwatches and wearable technologies have shown great prospect in detecting these conditions in a privacy-preserving and affordable way, compared to the conventional vision or text-based approaches.

According to previous studies [2][3], an individual's personal attributes (e.g., age, gender, history) can aid in more accurate detection of affect by helping the model learn discriminative features. Furthermore, affective events often have related contextual information that are significant [3]. For example, a person showing high electrodermal activity could be due to chronic stress history or due to an activity (e.g., running), or exposure to an environment (e.g., being outside). Although such contextual and personal attributes can be easily collected using smartwatches, extracting useful representations from these information and developing robust models for smartwatches remain a challenging job for several reasons. There is a dearth of expert-annotated datasets [4][5] to develop robust ML models. Often, a small portion of the data represent an affective event (e.g., a stressful situation), and thus making it difficult to capture the unique representation of the event. Moreover, without substantial amounts of labeled data, the models can easily learn irrelevant representations [4] from these additional personal and contextual information sources which could instead lead to performance degradation. As such, we need an efficient alignment strategy that allows the model to learn representation based on the contextual and personal similarities related to an event. Moreover, an affective event can have diverse characteristics depending on the domain, modality, context, and user. For example, the underlying representation of stress in one environment (e.g., the office) might be different from another (e.g., a hospital) [4]. Furthermore, the smartwatches often provide poor-quality signals compared to other devices. Therefore, the learning model should be able to extract robust features in each setting using these signals.

Recently, there is a surge of self-supervised ML models combined with the constrastive learning methods [6][7]. Such models learn the underlying representations from huge amounts of unlabeled samples based on their similarities and greatly reduces manual labeling effort. However, to date, the existing affect recognition works [7][8][9] have not explored these techniques to combine the contextual or personal representation with the physiological representation to improve the generalizability of the models.

In this paper, we overcome these notable limitations of the state-of-the-art. We develop *SPARC*, a multimodal ML framework for Person and Context based Affect Recognition using Smartwatches. The key contributions of this work are threefold: *First*, we present a novel systematic ML pipeline for robust affect detection using convenient smartwatches. It combines a multimodal self-supervised learning approach with an affect alignment strategy to improve the model's robustness using limited data. *Second,* unlike the existing works, *SPARC* integrates personal and contextual information along with the physiological representation to build a discriminative and improved representation of an affective event. *Third,* we conduct a real-life user study among 19 participants and show that *SPARC* substantially improves the existing models for three different affect domains. We extensively investigate the role of personal and contextual alignment as well as articulate the generalizibilty of *SPARC* in different settings to improve the state-of-the-art.

II. Architecture of *SPARC*

SPARC is a novel systemic ML pipeline for affect recognition (e.g., stress, anxiety, and emotion detection) using smartwatches. The architecture of *SPARC* (Fig. 1) consists of a novel affect alignment module and an efficient multimodal self-supervised contrastive learning module that combines physiological representation with the contextual and personal knowledge and thereby improves the accuracy and the robustness of the model. We describe the architecture and the training steps as follows:

Sirat Samyoun is with Cornell University, USA and John Stankovic is with University of Virginia, USA.

*Corresponding author: Sirat Samyoun (ss3938@cornell.edu)

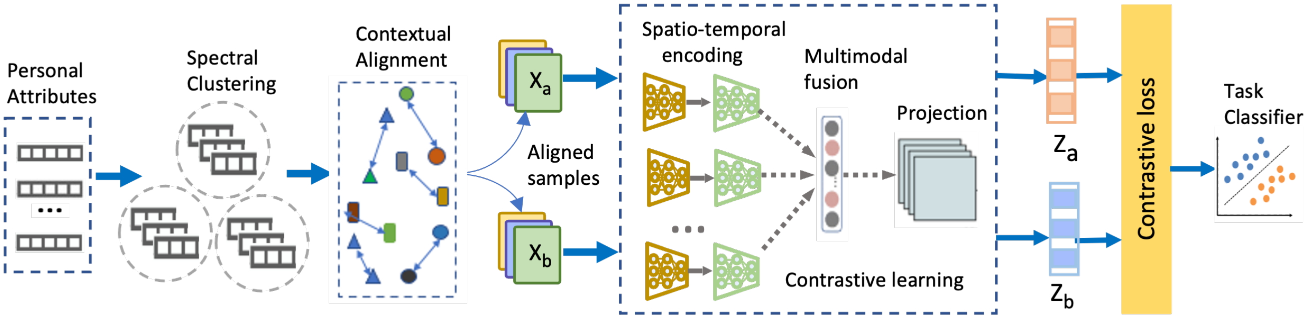

Fig. 1. The training steps of the *SPARC* framework for robust and accurate detection of affective states using personal and contextual information.

1) Step 1: Spectral clustering based person alignment: *SPARC* utilizes a spectral clustering method [10] to assign a profile group to a person based on individual attributes, such as, age, BMI, gender, and history. First, we aggregate the attributes for each person i as numerical values and normalize these values to form a person-specific vector, $p_i = [p_{i1}, p_{i2}, .., p_{im}] \in R^m$. Next, we build a fully connected similarity graph G along with the associated weighted adjacency matrix A. Two nodes i and j are connected in G, if $||p_i - p_j|| < \epsilon$, where $\epsilon > 0$, and ϵ is a tunable parameter. Next, we compute a degree matrix D, which is a diagonal matrix containing the degrees of each node as, $d_i = \sum_{j=1}^{N_s} w_{ij}$, where w_{ij} represents the edge between the nodes i and j in A, and N_s is the total number of users. We then build the Laplacian matrix for G as, $L = D - A$. Next, we normalize L and compute the Eigen vectors of L. We chose the first C Eigenvectors $(u_1, u_2, ..., u_c)$ and stack them into a matrix M with these vectors as the columns. Here, C is the number of desired profile groups, which is chosen by visual inspection. Next, we cluster these data points in M using a K-means clustering technique and form C clusters that represents our profile groups. Overall, this method ensures that the persons having similar profiles are aligned together to learn an invariant mapping across the cluster.

2) Step 2: Contextual alignment using pairwise similarity: To improve the model robustness, *SPARC* aligns the samples from a user based on their contextual similarities, such as, an inducer, an activity, and an environment [4]. To achieve this, *SPARC* first splits the data samples of each person into time windows. The contextual attributes for each window i are mapped into numerical values and then aggregated and normalized to form a contextual representation vector, $q_i = [q_{i1}, q_{i2}, .., q_{in}] \in R^n$. We hypothesize that each sample in an embedding space should have neighbors with similar contexts. To this end, we find out the contextual similarity of every window pair within a profile group by computing their cosine similarity. We then chose top-k positive anchors and the top-k negative anchors for each window by ranking their contextual similarities as: $[q_i^\intercal q_1, q_i^\intercal q_2, .., q_i^\intercal q_n], 1 \le i \le k$. Here, k is a parameter by design choice, and $\intercal$ computes the cosine similarity between their vectors. This strategy ensures that the samples that correspond to similar contexts maximizes the similarity in their learnt representations.

3) Step 3: Multimodal contrastive learning: Unlike the conventional approaches [7][8], *SPARC* extracts both spatial and temporal features from the data and applies multimodal fusion to achieve a comprehensive affective representation. It then applies a contrastive learning method, inspired by the SimCLR [6] framework. To further improve the performance, it uses the aforementioned aligned windows instead of requiring any random data augmentation. For each window, the top-k positive anchors are used to form $(k \times N)$ positive pairs, where N is the batch size. The equivalent negative samples are randomly chosen from temporally distant windows belonging to a different profile group or from the negative anchor group. Next, we apply the following steps:

Spatio-temporal encoding: To achieve a robust unimodal representation, *SPARC* captures both spatial and temporal correlations present in the data. First, it applies a CNN model (Φ) having 2 convolutional layers having 64 filters and Relu activation. It also has a maxpooling layer followed by two fully connected layers that provides the spatial representation of the ith sample for modality m as, $X_i^m = (x_{i1}^m, x_{i2}^m, \dots, x_{iN}^m)$, where $X = \Phi(W)$.

Next, we apply a LSTM model (Ψ) to the encoded spatial representation X_i^m. It splits the input into several time steps and passes it through a bidirectional LSTM layer with 100 cells and Relu activation function, followed by a fully connected layer. The bidirectional layers enable extracting the temporal representation in both directions, and thus we obtain the spatio-temporal representation $Y_i^m = (y_{i1}^m, y_{i2}^m, \dots, y_{iN}^m)$, where $Y = \Psi(X) = \Psi(\Phi(W))$.

Multimodal fusion: As opposed to learning from a particular modality [9], *SPARC* achieves a comprehensive understanding of an affective event by utilizing all modalities. This step concatenates the unimodal embedding vectors to form a multimodal representation, $Y_i = (Y_i^1 \oplus Y_i^2 \oplus .. \oplus Y_i^N)$.

Projection: The projection (Θ) starts with two convolutional layers having 32 and 16 filters respectively, each followed by a max-pooling layer having Relu activation. We used batch normalization to standardize the projected representations. Finally, we obtain the projected embedding for the sample W_i as, $Z_i = \Theta(Y_i) = \Theta(\Psi(\Phi(W_i)))$.

Next, we utilize a contrastive loss to maximize the agreement among positive pairs and train the framework in a self-supervised manner. For any pair (a, b) having projections Z_a and Z_b respectively, this loss is given by:

TABLE I

DETAILS OF DATA COLLECTED IN THE STUDY.

Data type	Description
Personal attributes	Age, gender, body-mass index, income
	History (*anxiety disorder, chronic stress*)
Contextual attributes	Inducer (*health/interpersonal/work/financial/other*)
	Activity (*walking/running/watching/eating/other*)
	Environment (*home/work/other*)
Physiological	BVP (64Hz), EDA (4Hz), Temperature(4Hz)
Affect labels	Stress, anxiety, emotional valence and arousal

$$L_{a,b} = -log(\frac{exp(sim(Z_a, Z_b)/\tau))}{\sum_{i=1}^{k.N} \mathbb{1}_{i \neq a} exp(sim(Z_a, Z_i)/\tau))})$$

Here, τ is the temperature parameter, $\mathbb{1}$ is the indicator function and $sim(.)$ is the cosine similarity, adopted from [6]. Overall, this step helps *SPARC* learn generalized representation based on contextual and personal similarities.

*4) **Step 4: Task-specific classification**:* We now utilize the learnt generalized representations by keeping the aforementioned modules frozen and then training a *Multi-layered Perceptron* model using the crossentropy loss with the target class labels. It has two fully connected layers and finally ends with a softmax function for affect classification. Thus, this step fine-tunes *SPARC* to learn task-specific representations.

Overall, this novel architecture combines affect related sensor readings with contextual and personal attributes utilizing multiple modalities and contrastive learning for a more comprehensive representation of an affective event.

III. EVALUATION

*A. **User study and experimental setup***

*1) **Study design**:* We designed and conducted a user study for collecting individuals affective data in real-life environments along with contextual and personal information using smartwatches. We experiment using this dataset only due to lack of publicly available smartwatch based affective datasets having relevant contextual and personal information that can be used to prove the claim of the paper. The study was approved by the University of Virginia IRB SBS Board (SBS#5234). It included 19 subjects, with 13 male and 6 female persons, all healthy individuals, aged 21-39 years (average 30 years). The participants were provided a smartwatch (Empatica E4 wristband) and a smartphone app (with a reminder app installed). The participants were asked to wear the watch throughout their daily activities. Each subject used the system for at least 2 weeks, and the overall study took around 5 months. Overall, four kinds of data were collected, such as, *personal attributes*, *contextual attributes*, *physiological data*, and *affect labels* using self-assessment surveys.

*2) **Data collection and processing**:* Table I provides a detailed overview of the collected data. The smartwatch recorded different physiological signals of the user: electrodermal activity (EDA), blood volume pulse (BVP) and skin temperature. During every 20 minutes, the smartphone app prompted the user for self-reporting affective states on a survey questionnaire. We chose 10 questions from the following scales used in the literature [4][5]: Positive and Negative Affect (PANAS) for stress labels, State-Trait Anxiety Inventory (STAI) for anxiety labels, Self-Assessment Manikins (SAM) for emotional valence and arousal labels. Similar to these works, the responses were recorded on a Likert scale of 1-5 and were converted to corresponding class labels, such as, stress (*neutral, stressed, amused*), anxiety (*high, low*) and emotions (*high and low-valence and arousal)*. If any of these states have changed during the last 20 minutes, the app further collected different contextual information from the user, such as, the inducer or source of the event, activity, and environment. The participants also provided their personal basic information and history related to stress or anxiety disorders. Following data collection, we followed the pre-processing techniques used in [4] and synchronised the physiological data with the labels.

TABLE II

PERFORMANCE FOR DIFFERENT ALIGNMENTS AND MODALITIES.

Task	Alignment				Modality			
	NA	CA	PA	*SPARC*	EDA	BVP	TEMP	*SPARC*
Stress	73.7	83.5	78.2	89.6	80.7	81.7	78.4	89.6
Anxiety	69.8	80.3	74.6	86.4	76.7	78.8	75.0	86.4
Emotion	71.2	81.6	77.3	87.1	77.6	79.5	71.2	87.1

*B. **Results and Performance Analysis***

*1) **Comparison across alignment techniques**:* We evaluate the impact of personal and contextual alignment in *SPARC* compared to the other baselines, these are:

- *Contextual Alignment (CA):* Only contextual alignment was applied. Step 1 was eliminated from the pipeline, so all users were put in a single group.
- *Personal Alignment (PA):* Step 2 was eliminated. Positive and negative pairs were chosen from the same profile group and different cluster, respectively.
- *No Alignment (NA):* No alignment technique was used.
- *SPARC:* The model was trained using all steps (1-4).

Table II presents the accuracy results for different classification tasks. Evidently, *SPARC* provided the best results, with an improved accuracy (by 6%-17%) across all tasks, and thus showing the impact of using affect alignment techniques. We also observe that the contextual information provided the most useful representation across all tasks.

*2) **Impact of sensing modalities**:* Table II also highlights the impact of different physiological modalities used in *SPARC*. Compared to the unimodal approaches, *SPARC* is able to learn general-purpose representation from all sensors and thus providing around (8%-11%) improvement. We also note that the modalities contributes differently in different tasks, with BVP being the most useful modality.

*3) **Robustness comparison under limited data settings**:* We evaluate the robustness of *SPARC* and several supervised models under different labeled data settings, such as:

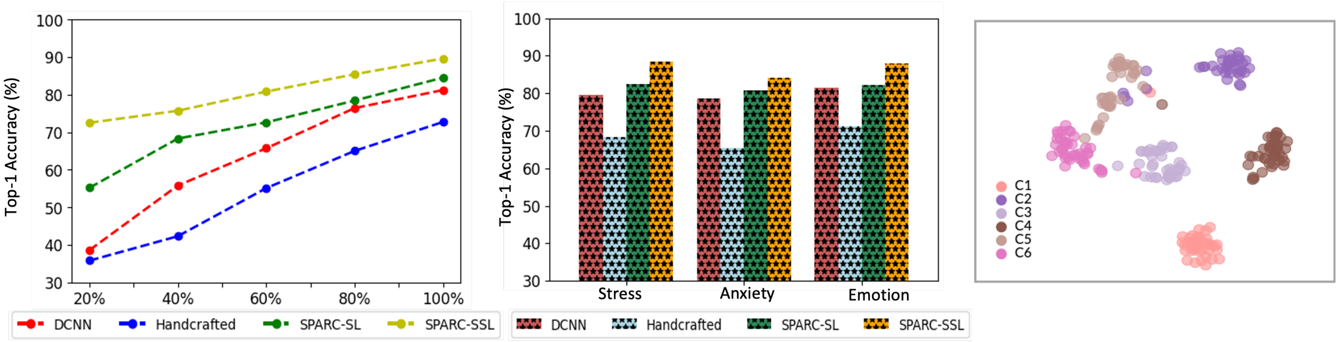

Fig. 2. (a) Robustness comparison with baselines under different data settings.(b) Generalizability comparison of *SPARC*. (c) t-SNE visualization

- *Handcrafted models*: We implement the feature-based models (e.g., *Random forest*) using the settings described in [4], and picked the best results.
- *Deep-CNN*: State-of-the-art popularly used CNNs [4][9] for affect recognition. We adopt the CNN structure from [9]. It has two convolutional layers, one maxpooling layer with dropout technique, and two dense layers with softmax activation for classification.
- *SPARC-SL*: Instead of the self-supervised training, end-to-end training was done in a supervised setting.
- *SPARC-SSL*: This implementation follows the original training architecture, with all steps 1-4.

Fig. 2(a) shows the accuracy results under different percentages of labeled data for the stress task, which depicts that *SPARC* substantially outperforms the baselines. We found similar results for the other tasks too. Particularly, with limited labels (20%-40%), *SPARC* provides reasonably good results (72%-75%), while the supervised models performs poorly. This will greatly help in developing robust affective models using low-frequency samples of a smartwatch, and also will reduce the need for manual affect labeling.

4) ***Generalizability to new users:*** We test the generalizability of *SPARC* to adapt to a new user compared to the baselines. In each round, data from all subjects were used for training except one, whose data was used for testing only. We repeat this for every testing subject and calculate the average accuracy across all rounds. The results (Fig 2(b)) shows that *SPARC* once again outperforms the baselines for all tasks. We believe this is because *SPARC*, by design can align a new user to a profile group and learns general-purpose representation from that group.

The results in Fig. 2(a)-2(b) also give the impression that the handcrafted models inherently can not capture any spatial or temporal representation from the limited smartwatch data well. While the DCNN and the *SPARC* variants overcomes this, the *SPARC*-SSL model excels among all by learning discriminative features that are based on contextual and personal information in a self-supervised manner.

5) ***Visualizing representation:*** We adopt t-distributed Stochastic Neighbor Embedding (t-SNE) to visually demonstrate the representation learned by *SPARC*. In this example, we used the self-supervised stress detection model and randomly generated 50 embeddings under a profile cluster and then assigned group labels (C1-C6) to the contexts. The emerged clusters (Fig. 2(c)) mostly shows clear boundaries which indicates that *SPARC* can form high-quality embeddings by utilizing personal and contextual attributes.

IV. Conclusion and Potential Impact

Overall, *SPARC* is capable of learning robust and discriminative affective representation by utilizing personal and contextual information. Compared to the existing works, *SPARC* is applicable for a range of domains and performs substantially well in limited data settings as well as for new users. Future works in research and industry can extend *SPARC* by designing and providing mental health interventions on commodity smartwatches.

V. Acknowledgements

We would like to thank Dr. Tanzeem Choudhury and Dr. Deborah Estrin from Cornell University, USA.

References

[1] “Anxiety disorders,” World Health Organization, https://www.who.int/news-room/fact-sheets/detail/anxiety-disorders.

[2] M. Stojchevska et al., “Assessing the added value of context during stress detection from wearable data,” BMC Med. Inform. Decis. Mak., vol. 22, no. 1, 2022.

[3] L. Abbruzzese, N. Magnani, I. H. Robertson, and M. Mancuso, “Age and gender differences in emotion recognition,” Front. Psychol., vol. 10, 2019.

[4] S. Samyoun, M. M. Islam, T. Iqbal, and J. Stankovic, “M3Sense: Affect-agnostic multitask representation learning using multimodal wearable sensors,” Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 6, no. 2, pp. 1–32, 2022.

[5] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, “Introducing WESAD, a multimodal dataset for wearable stress and affect detection,” in Proceedings of the 20th ACM International Conference on Multimodal Interaction, 2018.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in Proceedings of the 37th International Conference on Machine Learning, 2020, vol. 119, pp. 1597–1607.

[7] V. Dissanayake, S. Seneviratne, R. Rana, E. Wen, T. Kaluarachchi, and S. Nanayakkara, “SigRep: Toward robust wearable emotion recognition with contrastive representation learning,” IEEE Access, vol. 10, pp. 18105–18120, 2022.

[8] S. Samyoun, A. S. Mondol, and J. Stankovic, “A multimodal framework for robustly distinguishing among similar emotions using wearable sensors,” in 2022 44th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC), 2022.

[9] L. Shu et al., “A review of emotion recognition using physiological signals,” Sensors (Basel), vol. 18, no. 7, p. 2074, 2018.

[10] U. Luxburg. A tutorial on spectral clustering. Statistics and Computing, 17(4):395–416, 12, 2007.