A Multimodal Framework for Robustly Distinguishing among Similar Emotions using Wearable Sensors

Sirat Samyoun*, *Student Member, IEEE*, Abu Sayeed Mondol, and John Stankovic, *Fellow, IEEE*

Abstract— **Detecting the correct emotion is crucial for improved mental health outcomes. While the existing works on emotion recognition focus on detecting the common or primary emotions only, there exists some uncommon or secondary emotions of similar kind (e.g., *contempt* vs *anger*) that makes accurate emotion detection challenging. Moreover, there exists limited labeled data on such secondary emotions. We present the first work to accurately discriminate among such similar emotions by generating distinguishable data for the secondary emotions using convenient multimodal wrist sensors. Extensive evaluations show that our novel solution provides around 7-36% F1-score improvement to existing solutions for similar emotions, and also significantly reduces the burden on providing labeled emotion data.**

I. Introduction

Emotion detection is very important in mental health monitoring, and for choosing the correct interventions. People mostly exhibit common or primary forms of emotion in daily life, such as, *happy*, *angry*, *sad*, and *nervous*. However, there are emotions which are not so common, but are important to detect, such as, *contempt*, *disgust*,*frustration*, and *delight*. Interestingly, some of these uncommon or secondary emotions have very similar characteristics to the primary ones, and therefore are difficult to distinguish. For example, *contempt* is regarded as the biggest predictor of domestic violence, however, it is often confused with *anger* and *disgust* [1]. As another example, *frustration* and *confusion* have similarities with *nervous* [2], and such emotions are common among children at school [3]. Moreover, many similar negative emotions (e.g., *anger*, *frustration*, and *disgust*) were widespread among people since the COVID-19 pandemic [4].

Interestingly, such similar emotions can be uniquely represented using the Russel's valence-arousal emotion model [2]. For example, *delight* and *happy*, although being very similar, have unique locations in this model. However, despite the significant research effort over the years, existing works on emotion recognition [5-9] have mostly focused on accurately detecting primary emotions, and not focused on situations where such similar emotions are prevalent.

State-of-the-art works have also predominately used facial expression (captured via a video camera) or voice (via a microphone) [5-6]. However, these approaches are highly privacy-invasive, expensive, and non-ubiquitous for continuous monitoring. Moreover, people often conceal emotions with the facial expression or voice [7]. A suitable approach to overcome these limitations is to use physiological signals [7-9], such as, Electroencephalography (EEG), Electrodermal Activity (EDA), Blood Volume Pulse (BVP), Temperature (TEMP) from the wearables placed on the body (e.g., head, chest, and wrist). Among these, the wrist wearables or smartwatches provide the most privacy-preserving, convenient, and ubiquitous way to monitor emotions continuously.

Despite these benefits, there are several challenges to accurately distinguish among such similar emotions using wrist wearables. *First*, acquiring substantial amount of labeled data for both primary and secondary emotions is difficult, which limits the effectiveness of high-accuracy solutions (e.g., neural networks). *Second*, the available datasets are highly imbalanced, having abundant samples of only primary emotions. Moreover, straightforwardly using class distribution to augment new samples and balance the classes [10] does not guarantee better performance, particularly for secondary emotions that are highly similar to the primary ones. To solve this, we need to identify and generate the truly distinguishable samples of the secondary classes that do not follow the class distribution of primary ones. However, no prior works have done it. *Third*, the existing accuracy-based solutions [5-9] do not truly understand the difference among similar emotions. For example, an emotion classifier having abundant primary emotion samples much more than the secondary ones will show high accuracy even if it poorly detects the secondary class. Therefore, we need to build a proper solution that detects all classes well. *Fourth*, the wrist wearables provide very limited modalities having less accurate signals than the chest or head ones, which makes building an accurate multimodal solution challenging.

In this paper, we address these challenges and present the first work to accurately distinguish among such similar emotions. The key contributions of this paper are: **First**, we develop a novel framework, named *SEMSense* for accurately detecting all categories of emotion, specifically in presence of similar emotions, using the convenient wrist wearables only. We use a Convolutional neural network based model to generate robust multimodal representation of such similar emotions in the valence-arousal space. **Second,** we provide a new algorithm that generates distinguishable emotion class samples using class-conditional similarity estimation and thus overcomes the data imbalance problem. **Third,** our extensive experiments on a real-life dataset with different settings of parameters and modalities show that *SEMSense* provides around 7-36% F1-score improvement to the existing solutions, and also boosts the overall emotion detection performance in an all-in-one setting.

All the authors are with the Department of Computer Science, University of Virginia, Charlottesville, VA 22903, USA.

*Corresponding author: Sirat Samyoun (`ss8hf@virginia.edu`)

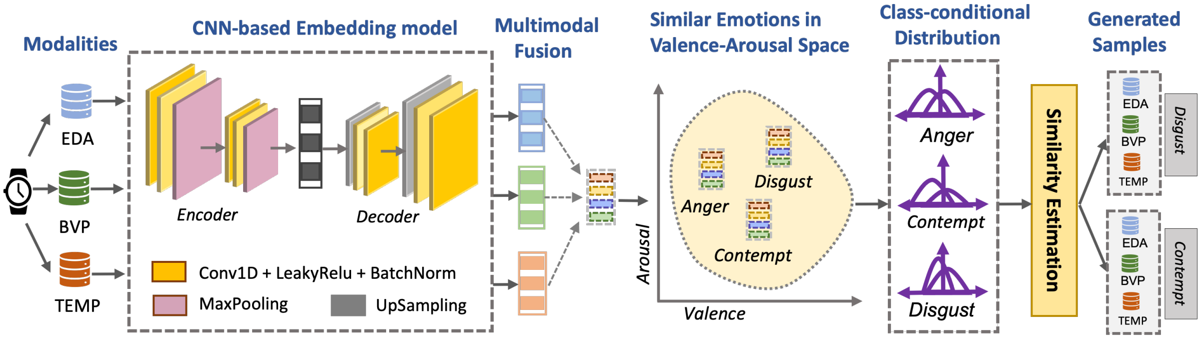


Fig. 1. Training steps of *SEMSense* framework for generating secondary class samples using original data from wrist modalities

II. Methodology of the Framework

SEMSense is a novel framework for producing new emotion samples that are distinguishable from other similar classes. The training procedure of *SEMSense* is shown in Fig. 1. It consists of two key steps. First, the original data samples are embedded in the valence-arousal space using an embedding model. Next, new secondary emotion samples are generated using class-conditional distribution of embeddings and then representative samples of the class are chosen based on their relative similarities with respect to the primary class. Following the training of *SEMSense*, the generated emotion samples are added to original training data for distinguishing among the similar primary and secondary classes.

A. Valence-Arousal Space Emotion Embedding

We develop a Convolutional neural network (CNN) based embedding model that efficiently projects data samples into a shared space that consists of valence and arousal dimensions. The model is an encoder-decoder network that effectively extracts the spatial correlations from the wrist data. First, the training data from a modality m is split into several sequences in the form $x_c^m = (s_c^m, v_c, a_c)$, where s_c^m is the raw data of class c, and v_i, a_i are the valence and arousal values of the sequence, respectively. Next, the sequences are passed through the encoder F_m and the decoder G_m. Given x_c^m, F_m extracts the spatial modality representation and produces an embedding vector $e_c^m \in R^d$ of dimension d. Similarly, given e_c^m, G_m aims to reconstruct the original sequence as $h_c^m \in R^n$. The model is trained end-to-end by optimizing the reconstruction loss $\mathcal{L}_{re}$, where θ_f^m and θ_g^m are the learned parameters of F_m and G_m, respectively.

$$\mathcal{L}_{re} = \|e_c^m - h_c^m\|_2^2$$
$$e_c^m = F_m\left(x_c^m; \theta_f^m\right), \; h_c^m = G_m\left(e_c^m; \theta_f^m\right)$$

1) Training details of the embedding model: The encoder starts with two 1D convolutional layers having 16 and 1 filters respectively, each followed by a max pooling layer of size 2 to produce the bottleneck embedding. The decoder applies two 1D convolutional layers having 1 and 16 filters respectively on the embedding, each followed by 2 sampling layers of size 2. Leaky Relu activation was used to introduce non-linearity, along with batch normalization to standardize the inputs. The model was trained with the Adam optimizer for utilizing an adaptive learning rate.

Algorithm 1: Embedding-based Emotion Generation

```
Input: $C$: Similar classes, $c_p$: Primary class, $c_q$:
       Secondary class, $r$: Augmentation factor, $e$:
       Embedding list, $M$: Available Modalities
Output: $L_q^m$: Generated representative samples of $c_q$.
for $c \in C$ do
    $e_c \leftarrow (e_c^1 \oplus e_c^2 \oplus e_c^3 \oplus ... \oplus e_c^m)$, for $m \in M$
    $(\mu_c, \Sigma_c) \leftarrow$ Compute using (1)
    $\mathcal{N}_c(Y_c|\mu_c, \Sigma_c) \leftarrow$ m.g.d using $(\mu_c, \Sigma_c)$
end
$K \leftarrow size(e) \times r\%$
while $size(L_q^m) < K$ do
    $e_{new} \leftarrow$ Sample from $\mathcal{N}_q(Y_q|\mu_q, \Sigma_q)$ for $c_q$
    $z_q \leftarrow MSS(e_{new}, \mathcal{N}_q)$ using (2)
    $z_p \leftarrow MSS(e_{new}, \mathcal{N}_p)$ using (2)
    if $z_q > z_p$ then
        $h_c^m \leftarrow G_m(e_{new})$, for $m \in M$
        $L_q^m \leftarrow L_q^m + \{h_c^m\}$
    end
end
```

2) Multimodal embedding fusion: Using the trained embedding model, we obtain unimodal spatial emotion embedding for each class. Next, we concatenate the embedding vectors from all modalities to generate a robust multimodal valence-arousal representation of the class, which is given by: $e_c = (e_c^1 \oplus e_c^2 \oplus ... \oplus e_c^m)$.

B. Emotion Sampling from Embedding Distribution

After generating the valence-arousal space representation, we model each similar emotion class using a class-conditional normal distribution. Specifically, for class c, we define a multivariate Gaussian distribution $\mathcal{N}_c(e_c|\mu_c, \Sigma_c)$. To obtain the distribution, we empirically compute the mean μ_c and covariance Σ_c from the training samples of count N_c:

$$\mu_c = \frac{1}{N_c}\sum_{i=1}^{N_c} e_{c,i}, \quad \Sigma_c = \tfrac{1}{N_c}\sum_{i=1}^{N_c}\left(e_{c,i} - \mu_c\right)\left(e_{c,i} - \mu_c\right)^\top \; (1)$$

TABLE I

PERFORMANCE COMPARISON FOR DISTINGUISHING AMONG SIMILAR EMOTIONS WITH AND WITHOUT *SEMSense*

Classification	**Experiment**	**F1-score(%)**					**Accuracy (%)**				
		Random Forest	Linear SVM	Extra Trees	Ada-Boost	Neural Network	Random Forest	Linear SVM	Extra Trees	Ada-Boost	Neural Network
Anger, Contempt, Disgust	Without *SEM*	49.3	29.1	42.7	43.9	35.7	82.4	78.3	83.9	68.1	74.9
	With *SEM*	**85.4**	56.1	74.0	60.4	57.9	**89.5**	87.2	88.3	69.5	87.1
Happy, Delight, Surprise	Without *SEM*	46.3	29.0	45.8	48.2	38.5	82.0	76.6	81.1	65.0	72.5
	With *SEM*	**57.7**	39.1	53.2	55.4	50.1	85.6	81.4	85.1	76.1	**87.2**
Nervous, Confusion, Frustration	Without *SEM*	38.1	31.7	38.7	25.4	40.6	91.1	90.4	91.0	48.2	87.6
	With *SEM*	56.1	49.2	54.4	47.7	**64.0**	91.5	91.0	91.6	84.3	**92.7**
All-in-one	Without *SEM*	52.9	43.4	51.8	57.2	49.9	77.3	72.5	77.2	68.7	70.1
	With *SEM*	**60.2**	53.1	58.4	60.1	58.1	**90.1**	87.3	90.3	83.4	89.2

1) Mahalanobis distance-based similarity estimation: Instead of using new samples directly from class distribution like existing works, *SEMSense* computes their class similarities first and then chooses the representative samples. Algorithm 1 depicts these steps, overall. We adopt Mahalanobis distance [11], a widely used metric for measuring the point's distance from a distribution. In a vector space, the similar embeddings will be less distance apart from each other. So, we draw a embedding e' from the secondary class distribution, and measure it's Similarity Score (MSS) with the related classes using the inverse Mahalanobis distance.

$$MSS(e', \mathcal{N}_c) = \{(e' - \mu_c)^T \Sigma^{-1} (e' - \mu_c)\}^{-1} \quad (2)$$

2) Representative secondary emotion samples selection: A drawn candidate having a higher similarity score with respect to the secondary class than the similar primary class is added to the representative samples set of the secondary class. Candidates are generated until the desired overall count is obtained according to a user-defined augmentation factor r. Next, the representative samples are fed through the trained decoder model for each modality to obtain the new raw samples, and are added to the original dataset.

III. EVALUATION

A. Experimental setup

1) Dataset Description: K-EmoCon [12] is a publicly available dataset for emotion detection using physiological signals that includes real-life data for both primary and secondary emotions. Data was collected from 32 participants, and were annotated by experts and participants themselves. It provides 3 wrist physiological signals collected by an Empatica E4 wristband: EDA, TEMP, and BVP. We develop the following classifiers for this dataset:

- *Anger, Contempt, Disgust:* These emotions are closely related in real-life [1], and in Russel's model [3].
- *Happy, Delight, Surprise:* These emotions are located in the positive valence and higher arousal region in [3].
- *Nervous, Confusion, Frustration:* These emotions are located in the negative valence higher arousal region.
- *All-in-one:* This classifier integrates all the aforementioned similar emotions. It will classify a sample to one of the 3 above groups, which will be next classified by an individual classifier of an emotion category.

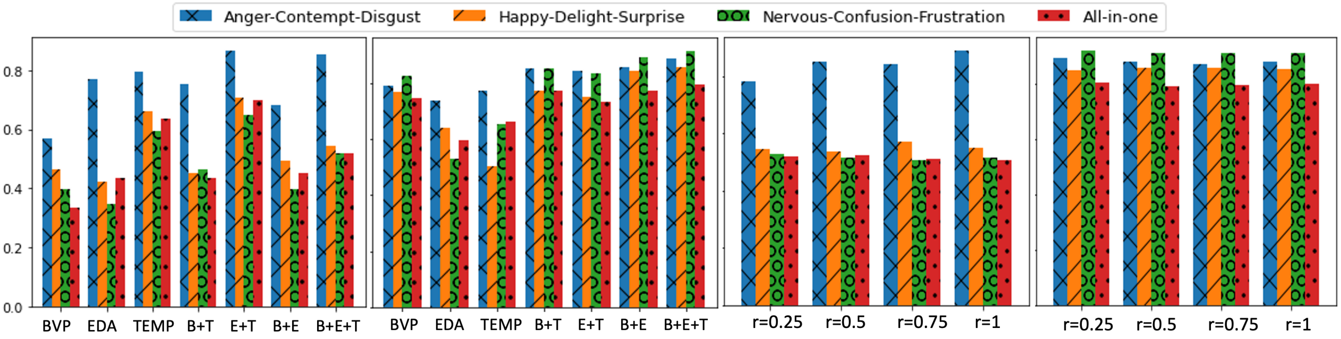
2) Pre-processing and feature extraction: First, the data samples for similar categories from all wrist modalities were merged with the aggregated annotations using the timestamp. The primary emotion annotations were converted to a scale of 0-1 to match with those of the secondary ones and the samples belonging to one category only were chosen. Next, the data were split into windows of 2 seconds and several statistical features (e.g., *Mean, Standard deviation, Minimum, Maximum, Skew, and Kurtosis*) were extracted.

3) Baselines and performance metrics: We apply *SEMSense* to all experimental cases by using different machine learning classifiers used in the state-of-the-art [5-9]. *Random Forest, Support Vector Machine, Extra-Trees, and Adaboost*, and an MLP-based *Neural Network* classifier were implemented using Scikit-learn library default settings. We measure the *Accuracy* and *F1-score* for evaluation.

B. Results and Performance Analysis

1) Comparison against baselines: Table 1 shows the classification results using *SEMSense* (with *SEM*) compared to that without using *SEMSense* in the pipeline (without *SEM*). For all experiments, our solution provided the best results (highlighted in bold), with significantly better F1-score (7%-36%) and accuracy (4%-24%) than the baselines. Moreover, the *Random Forest*, a tree-based method mostly produced superior F1-scores, because it chooses the best emotion features within the similar emotions from the underlying trees. We also note that regardless of the method, the F1-scores without *SEMSense* are mostly poor, which means detecting emotions becomes difficult in presence of similar primary and secondary emotions. Moreover, we used the wrist data only, which often provides less accurate signals. Using *SEMSense* greatly overcomes this problem, as the results justify, because it allows a classifier to better understand the decision boundary between the similar classes. Also, it is noteworthy that *SEMSense* provides good results in the all-in-one setting in terms of both F1-score and accuracy, which gives the impression that *SEMSense* can be used in any emotion detection pipeline that detects many emotions.

2) Accuracy vs F1-score for similar emotions: The accuracy results of most experiments are relatively high even if the corresponding F1-scores are low. For example, in *Nervous-Confusion-Frustration* classification, the solution without *SEMSense* produces very high accuracy (91.1%



(a) *Modality* vs *F1-score* (b) *Modality* vs *Accuracy* (c) *r* vs *F1-score* (d) *r* vs *Accuracy*

Fig. 2. Effect of modality combinations and augmentation factor (r) in *SEMSense* performance for detecting similar emotions.

for *Random Forest*), but much lower F1-score (38.1% for *Random Forest*). This happens because with imbalanced classes, even if a classifier poorly detects the secondary class, the accuracy will still be very high if the secondary class samples appear very few times in test set. This proves our claim that *SEMSense* would be more effective in detecting similar emotions for having better F1-score results.

3) Which modalities improve performance?: As modern smartwatches comes with different available modalities, it is essential to determine which modality combination works best for similar emotions. As per Algorithm 1, *SEMSense* can work with any combination of modalities. We experimentally show the effect of modalities in accuracy and F1-score performance for *SEMSense*, see Fig. 2(a) and 2(b), respectively. We observe that, a combination of EDA and TEMP modality works best in detecting the similar emotions. Our findings also suggest that using multimodal combinations over unimodal usually improve the performance. Moreover, we note that temperature (TEMP) is an important modality for distinguishing among such similar emotions.

4) Which value of r should be chosen?: We demonstrate the role of the augmentation factor r in *SEMSense* in the overall performance. Fig. 2(c) and 2(d) show the F1-score and accuracy results, respectively for $r = 0, 0.25, 0.75, 1$. All experiments used the *Random Forest* method, which provided best results in Table 1. We observe that increasing the value of r has very little effect on performance for most experiments. However, increasing the values of r means utilizing more labeled data using *SEMSense* during training whereas, the wrist devices have limited training resources. Our experiments suggest that choosing $r = 0.5$ provides the superior results, and therefore this value can essentially handle the trade-off between performance and training resources. Moreover, we see that all values of r provide consistently good accuracy performance for all similar emotions.

IV. Conclusion and Future Works

In this paper, we took the first step in the literature towards accurately distinguishing between the primary and secondary emotions of similar kind by utilizing data from the convenient wrist devices only. As the COVID-19 pandemic has triggered significant emotional contagion and disorders among people, smartwatches, using *SEMSense* can play a vital role by facilitating early diagnosis of such disorders and conveying further interventions to the user. Conducting such user studies is a long-term goal of this work. Moreover, the labeled data generated by *SEMSense* can help to minimize the burden on expert-annotated data, which poses a significant challenge to emotion recognition research. Finally, *SEMSense* provides a generic and scalable platform that can integrate any number of modalities and emotions, which will be usable with future smartwatches with more physiological features. Another future extension is to generate emotion data based on subjective differences for user personalization.

References

[1] J. Sommer, S. Iyican, and J. Babcock, The relation between contempt, anger, and intimate partner violence:A dyadic approach, J. of interpersonal violence,vol. 34, no. 15, pp. 3059–3079, 2019.

[2] J. A. Russell, A circumplex model of affect. J. of personality and social psychology, vol. 39, no. 6, p.1161, 1980.

[3] P. Watson, T. J. Ryan, Duration of the frustration effect in children. J. of Experimental Child Psychology, 4(3), pp.242-247, 1966.

[4] N. Montemurro, The emotional impact of COVID-19: From medical staff to common people. Brain, behavior, and immunity, 2020.

[5] M. Mansoorizadeh and N. M. Charkari, Multimodal information fusion application to human emotion recognition from face and speech, Multimedia Tools and Applications, vol. 49, no. 2, pp. 277–297, 2010.

[6] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, Analysis of eeg signals and facial expressions for continuous emotion detection, IEEE Trans. on Affective Computing, vol. 7, no. 1, pp. 17–28, 2015.

[7] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, 2018. A review of emotion recognition using physiological signals. Sensors, 18(7), p.2074.

[8] M. Ragot, N. Martin, S. Em, N. Pallamin, and J. Diverrez. Emotion recognition using physiological signals: laboratory vs. wearable sensors. In Int. Conf. on Applied Human Factors and Ergonomics. Springer, 15–22, 2017.

[9] J. C. Quiroz, E. Geangu, and M. H. Yong. Emotion recognition using smart watch sensor data: Mixed-design study. JMIR mental health, 5(3), e10153, 2018.

[10] X. Zhu, Y. Liu, J. Li, T. Wan, and Z. Qin. Emotion classification with data augmentation using generative adversarial networks. In Pacific-Asia Conf. on knowledge discovery and data mining. Springer, 349–360, 2018.

[11] P. C. Mahalanobis. On the generalised distance in statistics. Proc. of the National Institute of Sciences of India. 2 (1): 49–55, 1936.

[12] C. Y. Park, N. Cha, S. Kang, A. Kim, A. H. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, Kemocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations, Scientific Data, vol. 7, no. 1, p. 293, 2020.