2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

A Comparative Study of Arousal and Valence Dimensional Variations for Emotion Recognition Using Peripheral Physiological Signals Acquired from Wearable Sensors*

Feryal A. Alskafi, Ahsan H. Khandoker, *Senior Member IEEE*, and Herbert F. Jelinek, *Member, IEEE*

Abstract— **Wearable sensors have made an impact on healthcare and medicine by enabling out-of-clinic health monitoring and prediction of pathological events. Further advancements made in the analysis of multimodal signals have been in emotion recognition which utilizes peripheral physiological signals captured by sensors in wearable devices. There is no universally accepted emotion model, though multidimensional methods are often used, the most popular of which is the two-dimensional Russell's model based on arousal and valence. Arousal and valence values are discrete, usually being either binary with low and high labels along each dimension creating four quadrants or 3-valued with low, neutral, and high labels. In day-to-day life, the neutral emotion class is the most dominant leaving emotion datasets with the inherent problem of class imbalance. In this study, we show how the choice of values in the two–dimensional model affects the emotion recognition using multiple machine learning algorithms. Binary classification resulted in an accuracy of 87.2% for arousal and up to 89.5% for valence. Maximal 3-class classification accuracy was 80.9% for arousal and 81.1% for valence. For the joined classification of arousal and valence, the four-quadrant model reached 87.8%, while the nine-class model had an accuracy of 75.8%. This study can be used as a basis for further research into feature extraction for better overall classification performance.**

I. Introduction

Emotions play a vital part in human behavior and psychology, exerting a powerful influence on processes such as perception, attention, decision-making, learning, and general well-being. To understand human nature, cognition and intellect, investigating and classifying emotional states is important [1]. Emotions are usually classified as levels of arousal and valence with high or low arousal and valence indicating possible emotional dysregulation. In healthcare, the opportunity to build an individual profile that recognizes sources of stress, anxiety, depression, or chronic diseases can be achieved by wellness tracking and possibly including emotion recognition and classification determined from wearable apps [2]. The automatic assessment of emotional states can then assist in developing treatment protocols for mental and physical disorders [3].

Emotions are usually conveyed through body language, which can include facial expressions, body gestures, and intonation of voice. However, physiological manifestations of emotions can provide a more accurate representation since they occur subconsciously, are much harder to conceal, and are more difficult to manipulate compared to body language [3]. Not only that, but some disorders, such as autism, present impairments in facial expression, body postures, and movements associated with social interactions as well as deficient or deviant reaction to the emotions of other people [4]. Expression of emotions is linked with autonomic nervous system (ANS) activity, which in turn controls heart rate (HR), electrodermal activity (EDA), temperature and respiration patterns that can be used for determining emotion [3]. For this purpose, the emphasis in this research was on the identification of human arousal and valence states based on peripheral physiological signals acquired from wearable devices.

II. Experimental Method

A. Data

The K-EmoCon dataset with comprehensive annotations of continuous emotions during naturalistic conversations was chosen to conduct the investigation. The dataset contains multimodal measures recorded during 16 sessions of 10-minute paired debates between male and female students (age: 19 to 36), on a social topic. It contains emotion annotations from self, debate partner, and outside observers [5]. K-EmoCon provides data on emotion recognition outside a controlled laboratory condition. Blood volume pulse (BVP), electrodermal activity (EDA), heart rate (HR), and temperature, captured by the *Empatica E4 Wristband,* coupled with the self-annotations were used for emotion recognition. The *Empatica E4* has been previously validated for use in emotion recognition [6].

B. Pre-processing

BVP and HR signals were to the EDA and temperature signals by resampling to 4Hz. The signals were then normalized to a range between -1 and 1.

Information from 5-second segments that match the emotion labels collected in the dataset was then extracted for analysis. Each segment includes both arousal and valence, annotated by the participants on a scale of 1 to 5. For binary dimensions, 1 and 2 are considered low while 3, 4, and 5 are considered high. For the 3 class dimensions, scores of 1 and 2 are considered low, 3 is neutral, and 4 and 5 are high. Table 1 shows the total number of labels available for each score.

* This work was supported by a KU-KAIST joint grant (8474000221;KKJRC-2019-Health2 (Khandoker)) from Khalifa University Abu Dhabi, UAE.

F. A. Alskafi is with the Department of Biomedical Engineering, Khalifa University, Abu Dhabi 127788, UAE. (e-mail: 100057997@ku.ac.ae).

A. H. Khandoker and H. F. Jelinek are with the Department of Biomedical Engineering and the Health Engineering Innovation Center, Khalifa University, Abu Dhabi 127788, UAE. (e-mail: ahsan.khandoker@ku.ac.ae; Herbert.jelinek@ku.ac.ae).

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

Table 1. Total number of available arousal and valence labels for each score

Score	Arousal	Valence
1	148	72
2	933	483
3	1193	1900
4	702	752
5	323	92

Fig.1. shows a 1-minute excerpt (12 segments) from the signals for one of the participants after preprocessing with the corresponding arousal labels.

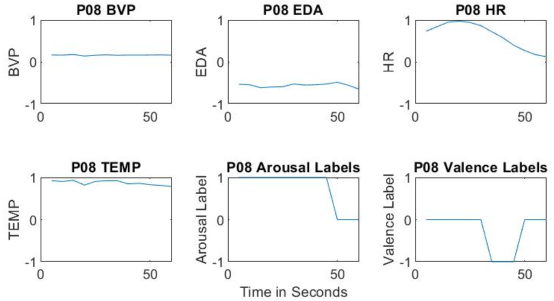


Figure 1. BVP, EDA, HR, and Temperature signals for participant 8. The low, neutral, and high classes correspond to -1, 0, and 1 respectively for arousal and valence labels.

C. Classification Models and Platform

The experimental setup included the implementation of machine-learning algorithms using the MATLAB classification learner application with its preset parameters. Decision trees (fine, medium, and coarse), support vector machines (linear, quadratic, cubic, fine Gaussian, medium Gaussian, and coarse Gaussian), k-nearest neighbors (fine, medium, coarse, cosine, cubic, and weighted), kernel naive Bayes, and ensembles (boosted trees, bagged trees, subspace discriminant, subspace KNN, and RUSBoosted Trees) classifiers were trained using the preset hyperparameters as a starting point. A holdout validation scheme was used with 30% of data held out for testing.

III. Results

Fig. 2. displays the results of comparing the performance of multiple classifiers for the different dimensionality variations.

A. Binary Classification

For the classification of 324 low and 665 high arousal instances, five classifiers achieved accuracy above 80% (Table 2).

Table 2. Accuracy and correctly classified instances for binary classification of arousal.

Classifier	Accuracy %	Correctly Classified Low Instances %	Correctly Classified High Instances %
Fine Gaussian SVM	84.6	67	**93.2**
Fine KNN	**87.2**	**79**	91.1
Medium KNN	81.5	67.3	88.4
Cubic KNN	80.9	66.4	88
Weighted KNN	86.8	74.4	92.8

Maximum values in each column are in bold.

The classification of 166 low and 823 high valence instances resulted in six classifiers achieving accuracy above 85% (Table 3).

Table 3. Accuracy and correctly classified instances for binary classification of valence.

Classifier	Accuracy %	Correctly Classified Low Instances %	Correctly Classified High Instances %
Fine Gaussian SVM	**89.5**	52.4	97
Fine KNN	88	**69.9**	91.6
Medium KNN	87.4	48.2	95.3
Cubic KNN	87.4	47	95.5
Weighted KNN	89.1	62.7	94.4
Ensemble Bagged Trees	87.8	39.8	**97.4**

B. Three-Class Classification

The classification of 324 low, 357 neutral, and 308 high arousal instances resulted in six classifiers achieving accuracy above 70% (Table 4).

Table 4. Accuracy and correctly classified instances for 3-class classification of arousal.

Classifier	Accuracy %	Correctly Classified Low Instances %	Correctly Classified Neutral Instances %	Correctly Classified High Instances %
Fine Gaussian SVM	79.9	**81.5**	**81.2**	76.5
Fine KNN	**80.9**	81.2	79.8	**81.8**
Medium KNN	71.1	79.1	68.9	65.1
Cubic KNN	70.7	76.6	70.9	64.2
Weighted KNN	80	80	79.3	80.8
Ensemble Bagged Trees	74.9	73.8	76.2	74.6

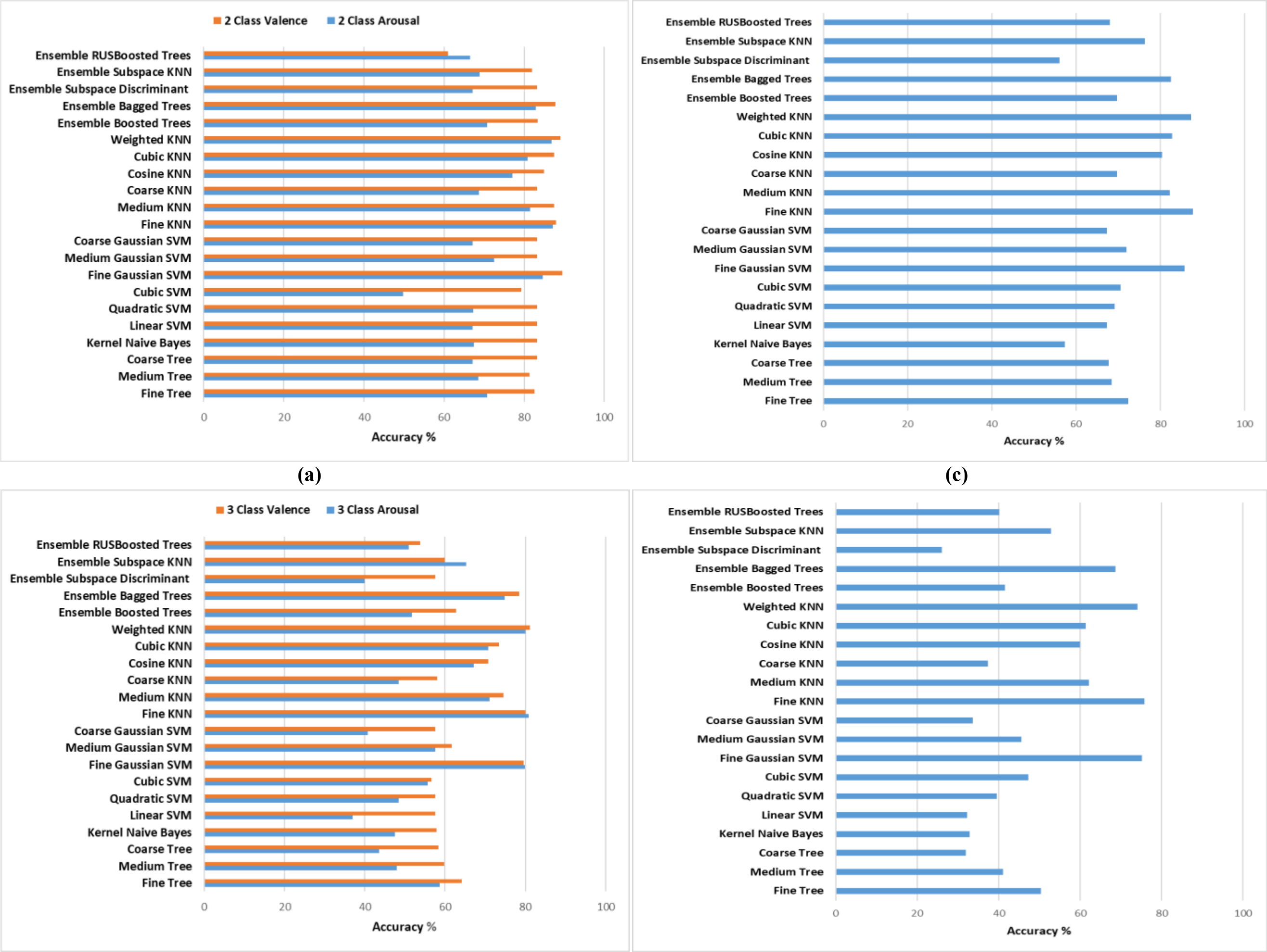
For Valence, the classification of 166 low, 570 neutral, and 253 high valence instances resulted in seven classifiers achieving accuracy above 70% (Table 5).

Table 5. Accuracy and correctly classified instances for 3-class classification of valence.

Classifier	Accuracy %	Correctly Classified Low Instances %	Correctly Classified Neutral Instances %	Correctly Classified High Instances %
Fine Gaussian SVM	79.5	50.6	**92.8**	68.4
Fine KNN	80	**69.9**	84.9	75.5
Medium KNN	74.5	43.3	89.8	60.5
Cosine KNN	70.7	41	86.8	53.8
Cubic KNN	73.5	42.2	88.6	60.1
Weighted KNN	**81.1**	56.6	90.2	**76.7**
Ensemble Bagged Trees	78.5	59	91.2	62.5

C. Quadrant Classification

The classification of 51 low valence/low arousal, 550 high valence/high arousal, 115 low valence/high arousal, and 273 high valence/low arousal instances resulted in seven classifiers achieving accuracy above 80% (Table 6).



(a)

(c)

(b)

(d)

Figure 2. Comparison of classifiers for the classification of arousal and valence using BVP, EDA, HR, and Temperature signals. (a) Arousal and valence classified into low and high each. (b) Arousal and valence are classified into low, neutral, and high each. (c) Arousal and valence jointly classified into four classes: LALV, HAHV, LAHV, and HALV (d) Arousal and valence jointly classified into nine classes: LVHA, NVHA, HVHA, LVNA, NVNA, HVNA, LVLA, NVLA, and HVLA.

Table 6. Accuracy and correctly classified instances for the joined 4-class classification of arousal and valence.

Classifier	**Accuracy %**	**Correctly Classified LALV Instances %**	**Correctly Classified HAHV Instances %**	**Correctly Classified LAHV Instances %**	**Correctly Classified HALV Instances %**
Fine Gaussian SVM	85.8	64.7	92.9	71.3	**95.7**
Fine KNN	**87.8**	**76.5**	92.4	**77.6**	94.8
Medium KNN	82.3	52.9	88.2	71	93.9
Cosine KNN	80.4	52.9	86	68.4	93.9
Cubic KNN	82.8	52.9	88.7	71.7	93.9
Weighted KNN	87.3	70.6	**93.8**	74.6	93
Ensemble Bagged Trees	82.5	51	91.1	66.5	93

LALV: Low Arousal/Low Valence, HAHV: High Arousal/High Valence, LAHV: Low arousal/High Valence, HALV: High Arousal/Low Valence .

D. Nine-Class Classification

Table 7 shows the results of seven classifiers with an above 60% when classifying 52 low valence/low arousal, 49 low valence/high arousal, 65 low valence/neutral arousal, 95 high valence/high arousal, 85 high valence/low arousal, 73 high valence/neutral arousal, 188 neutral valence/low arousal, 219 neutral valence/neutral arousal, and 163 neutral valence/high arousal instances.

IV. Discussion

A. Binary Classification

In the binary classification models, the results were in agreement with previous studies that used the same binary dimensions and machine learning classifiers [7]. The higher percentage of correctly classified high instances is due to the class imbalance [8]. Fine KNN performed best for the arousal model in terms of both accuracy at 87.2% and percentage of correctly classified low instance at 79%. For valence, the overall accuracy was higher, but the performance in regard to the low-class classification was lower. Fine Gaussian SVM had the highest accuracy at 89.5% though it was third-best in terms of correctly classified low instances at 52.4% compared to Fine KNN at 69.9%.

B. Three-Class Classification

Since we are generally interested in identifying negative emotion, applying the neutral label to the low would establish a false balance. The neutral emotion condition is both the most dominant and the most uncertain emotion form of most everyday situations [9]. The overall accuracy decreased for both models but remained in line with expected values [7]. The percentage of correctly classified low instances increased in the arousal model but decreased in the valence. For both models, it can be inferred that Fine KNN is the suitable choice.

C. Quadrant Classification

The four quadrants were defined according to the combined level of arousal and valence representing the emotion circumplex. In the current analysis, the low valence/low arousal class has the least number of samples compared to the high valence/high arousal class. Though the overall accuracy is high, LALV and HAHV percentages reflect the imbalance. Once again Fine KNN is the best classifier at 87.8% accuracy and 76.5% correctly classified LALV instances.

D. Nine-Class Classification

The number of samples per class for this model shows data imbalance around neutral valence, less so with arousal. Overall accuracy is lower, with Fine KNN the highest at 75.8%, but the deciding factor is the percentage of correctly classified instances. Fine Gaussian SVM, Fine KNN, and Weighted KNN, all perform best for two low-occurring classes, leaving the decision to the highest accuracy with Fine KNN.

V. Conclusion

BVP, EDA, HR, and temperature were used as predictors for multiple classifiers trained using variations of emotion dimensions and class distinctions. Performance was observed according to both accuracy and number of correctly classified instances as assessing a model's performance by accuracy alone is not useful in such applications where the datasets are class imbalanced and biased towards certain classes. The choice of classifier must take into consideration the low-occurring classes as they are usually the target of classification optimization. Overall, the Fine KNN classifier was found to perform best with this current dataset. The various dimensions showcased that higher number of classes decreased the overall emotion recognition accuracy. For the separate arousal and valence models, the 2-class models performed better accuracy-wise while the 3-class models showed higher balance between classes in terms of correctly classified instances. The same can be observed for the joined models.

Table 7. Accuracy and correctly classified instances for the joined 9-class classification of arousal and valence.

Classifier		**Fine Gaussian SVM**	**Fine KNN**	**Medium KNN**	**Cubic KNN**	**Weighted KNN**	**Ensemble Bagged Trees**
Accuracy %		75.1	**75.8**	62.1	61.4	74	68.6
Correctly Classified Instances %	**LVHA**	**85.7**	83.7	71.4	67.3	83.7	75.7
	NVHA	72.2	**77.2**	64.8	66.7	71.6	71.6
	HVHA	57.9	67.4	49.5	49.5	**69.5**	48.4
	LVNA	**84.8**	81.8	65.2	66.7	80.3	80.3
	NVNA	**82.7**	75.9	71.4	71.4	76.8	70
	HVNA	59.5	**66.2**	48.6	40.5	56.8	54.1
	LVLA	86.3	**94.1**	68.6	62.7	90.2	78.4
	NVLA	**78.6**	76.5	59.9	59.9	73.8	72.7
	HVLA	65.9	69.4	51.8	51.8	**71.8**	65.9

LVHA: Low Valence/High Arousal, NVHA: Neutral Valence/High Arousal, HVHA: High Valence/High Arousal, LVNA: Low Valence/Neutral Arousal, NVNA: Neutral Valence/Neutral Arousal, HVNA: High Valence/Neutral Arousal, LVLA: Low Valence/Low Arousal, NVLA: Neutral Valence/Low Arousal, HVLA: High Valence/Low Arousal.

Acknowledgment

This work was supported by a KU-KAIST joint grant (8474000221;KKJRC-2019-Health2 (Khandoker)) from Khalifa University Abu Dhabi, UAE.

References

[1] R. A. Calvo and S. D'Mello, "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications," in *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18-37, Jan. 2010, doi: 10.1109/T-AFFC.2010.1.

[2] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion Recognition from Multimodal Physiological Signals for Emotion Aware Healthcare Systems," *Journal of Medical and Biological Engineering*, vol. 40, no. 2, pp. 149–157, 2020.

[3] M. Egger, M. Ley, and S. Hanke, "Emotion Recognition from Physiological Signal Analysis: A Review," *Electronic Notes in Theoretical Computer Science*, vol. 343, pp. 35–55, 2019.

[4] M. Taj-Eldin, C. Ryan, B. O'Flynn and P. Galvin , "A Review of Wearable Solutions for Physiological and Emotional Monitoring for Use by People with Autism Spectrum Disorder and Their Caregivers," *Sensors,* 2018.

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

[6] M. Ragot, N. Martin, S. Em, N. Pallamin, and J.-M. Diverrez, "Emotion Recognition Using Physiological Signals: Laboratory vs. Wearable Sensors," *Advances in Human Factors in Wearable Technologies and Game Design*, pp. 15–22, 2017.

[7] P. Schmidt, A. Reiss, R. Dürichen, and K. V. Laerhoven, "Wearable-Based Affect Recognition—A Review," *Sensors*, vol. 19, no. 19, p. 4079, 2019.

[8] X. Guo, Y. Yin, C. Dong, G. Yang and G. Zhou, "On the Class Imbalance Problem," *2008 Fourth International Conference on Natural Computation*, 2008, pp. 192-201, doi: 10.1109/ICNC.2008.871.

[9] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.