Original scientific paper

Comparison of Machine Learning Approaches to Emotion Recognition Based on DEAP Database Physiological Signals

Tamara Stajić, Jelena Jovanović, Nebojša Jovanović, and Milica Janković

***Abstract* — Recognizing and accurately classifying human emotion is a complex and challenging task. Recently, great attention has been paid to the emotion recognition methods using three different approaches: based on non-physiological signals (like speech and facial expression), based on physiological signals, or based on hybrid approaches. Non-physiological signals are easily controlled by the individual, so these approaches have downsides in real world applications. In this paper, an approach based on physiological signals which cannot be willingly influenced (electroencephalogram, heartrate, respiration, galvanic skin response, electromyography, body temperature) is presented. A publicly available DEAP database was used for the binary classification (high vs low for various threshold values) considering four frequently used emotional parameters (arousal, valence, liking and dominance). We have extracted 1490 features from the dataset, analyzed their predictive value for each emotion parameter and compared three different classification approaches – Support Vector Machine, Boosting algorithms and Artificial Neural Networks.**

***Keywords* — DEAP database, emotion recognition, machine learning, physiological signals.**

I. Introduction

EMOTION is a complex behavioral phenomenon which includes different levels of neural activations and chemical reactions in the human brain [1]. Emotion is a combination of human thought, feeling and behavior, and can be defined as a physiological reaction to different external stimuli [2].

Paper received June 23, 2022; accepted October 27, 2022. Date of publication December 26, 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Branimir Reljin.

This paper is revised and expanded version of the paper presented at the 29th Telecommunications Forum TELFOR 2022 [14].

This research was supported by the Ministry for Education, Science and Technology Development of the Republic of Serbia, Belgrade, Serbia (contract 451-03-68/2022-14/200103).

Tamara Stajić is with the School of Electrical Engineering, University of Belgrade, Serbia (e-mail: tasa.stajic@gmail.com).

Jelena Jovanović is with the School of Electrical Engineering, University of Belgrade, Serbia (e-mail: jelenajovanovic0119@gmail.com).

Nebojša Jovanović is with the School of Electrical Engineering, University of Belgrade, Serbia (e-mail: nebojsa.php@gmail.com).

Milica Janković is with the School of Electrical Engineering, University of Belgrade, Serbia (e-mail: piperski@etf.rs).

For decades, emotions and emotion recognition have attracted a lot of attention which resulted in a variety of approaches that could be grouped into two distinct categories. The first group consists of methods based on non-physiological data such as speech [3] and facial expressions [4]. The advantage of this approach is the fact that the data is easily collected, without the need for any specialized and costly equipment. However, non-physiological signals can be willingly controlled which means that individuals can mask their emotion, and cause uncertainty in the classification that cannot be detected and removed. The second group relies on physiological data such as electroencephalography (EEG) [2], electromyography (EMG) [5], electrocardiography (ECG) [6], galvanic skin response (GSR) [7], etc. This approach allows better correlation with an actual emotional state, but at the same time makes it harder to set up the experiment, requires special equipment and subject preparation. Noise inherently present in these signals can also present an obstacle to reliable emotion recognition.

Hybrid approaches imply multimodal methods for emotion recognition that combine non-physiological and physiological approaches. Huang et al. [8] proposed a combination of facial expressions and EEG signals for emotion recognition, Fig. 1. The same approach was used by Tan et al. [9]. A Python package for the same task called `MindLink-Eumpy` was introduced by Li et al. [10]. In theory, this allows taking the best of both methods which should result in a higher accuracy.

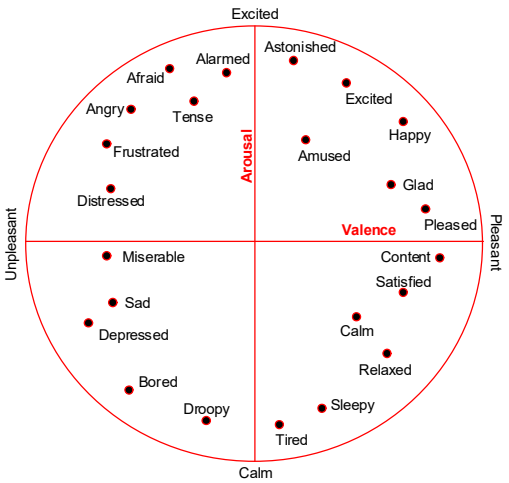

Fig. 1. The arousal/valence space.

The most common dimensional space used for describing emotions is the arousal/valence space, where emotions are described in terms of the intensity - going from 'inactive' to 'active' in the arousal dimension, and from 'unpleasant' to 'pleasant' in the valence dimension [11].

Aside from valence and arousal, other parameters commonly used in literature to present emotions are dominance (ranging from 'helpless' to 'in control') and liking.

These four parameters, alongside familiarity, are also used in the most used open database of physiological signals for emotion classification - DEAP [12].

The reported accuracies in the paper that has introduced the DEAP database [12] are 65.1%, 62.7% and 67.7%, and the F1 score 61.8%, 60.8% and 63.4% for arousal, valence, and liking, respectively. Torres-Valencia et al. [11] reported the highest accuracy value of 75% for binary arousal classification (high vs. low) when combining EEG, GSR, and ECG signals, and 58.7% accuracy for binary valence classification in case of using only EEG signals (F1 scores were not reported). Yang et al. [13] opted for an approach using a multi-column CNN-based model using EEG signals and reported accuracies of 90% and 90.6% for valence and arousal respectively.

An important aspect of emotion recognition is subjectivity. Emotion itself is a subjective occurrence which makes it difficult to generalize. There are two approaches regarding this issue – inter-subject and cross-subject. Even though inter-subject classification gives a higher accuracy in general, its applicability in real world cases is limited because it requires model recalibration or retraining for each new user, which can be very costly and time consuming. In this study, the cross-subject approach was chosen due to higher real-world applicability.

This paper provides an extended and modified version of results first presented in the paper "Emotion Recognition Based on DEAP Database Physiological Signals" [14]. The main goal of the study is a broad analysis of all available physiological data from the DEAP database, as well as evaluation of different machine learning algorithms for the purpose of emotion recognition. In this paper, we present an improved feature selection method, and a more detailed analysis of results together with the effects of choosing different thresholds for binary class definitions.

II. Methodology

A. The DEAP database

The DEAP database consists of 40 physiological signals from 32 subjects recorded while watching 40 different music videos. After each video, the subjects gave ratings, based on which emotions are labelled. The physiological signals included in the dataset are: 32-ch electroencephalogram (EEG), 2-ch electrooculogram (EOG), plethysmogram, respiration pattern, 2-ch electromyography (on zygomaticus major muscle, zEMG and trapezius muscle, tEMG), galvanic skin response (GSR) and body temperature. The sampling rate for all signals was set to 512 Hz. The DEAP database also includes recordings of facial expressions, which were not considered in our research.

In this paper, we used the pre-processed data available at https://www.eecs.qmul.ac.uk/mmv/datasets/deap/.

The parsed, pre-processed, and down sampled (by factor 4) data has the dimension 40x40x8064 which correspond to (number of videos) x (number of physiological channels) x (number of samples in one recording).

B. Signal processing and feature extraction

Data analysis was done in the Python programming language (Python Software Foundation, Delaware, USA). Aside from standard libraries used for scientific analysis like `NumPy` [15] and `SciPy` [16], we used the `pyphysio` library [17] for signal processing and feature extraction. The complete project code and further information is available at the following GitHub repository: https://github.com/nebojsa55/EmotionRecognition.

A review of all generated features (extracted from 8064 samples for each subject and for each video) is given in Table 1.

The focus of EEG analysis was on the statistical features of different frequency bands such as alpha (8-12 Hz), beta (13-30 Hz), gamma (30+Hz), and theta (4-8 Hz). Other EEG features included power spectral density (PSD) in different bands and Hjorth features (activity, mobility, and complexity) [18]. The resulting set consists of 44 features for each of the 32 EEG channels.

Respiratory signal features were extracted in the same way as features for heart rate variability - using the `hrvanalysis` [19] Python library. Before the analysis, the signal was filtered using a low-pass Butterworth filter (order 2, f_{low}=32 Hz).

Galvanic skin response (GSR) signal has two basic components – DC component which represents general activity of the sweat glands, and skin conductance response (SCR) component that is a good indicator of arousal level due to external sensory and cognitive stimuli [20]. A low-frequency drift was extracted from the GSR signal by applying a Moving Average (MA) filter, which was then subtracted from the GSR signal. This way the SCR component was singled out and additionally filtered by low pass (LF) fir filter (f_{low}=0.2 Hz) to obtain LF SCR signal and by very low pass (VLF) fir filter (f_{low}=0.08 Hz) to obtain VLF SCR.

EMG features were extracted from raw tEMG and zEMG signals, low pass fir filtered tEMG and zEMG signals (f_{low}=0.3 Hz) and very low pass fir filtered tEMG and zEMG signals (f_{low}=0.08 Hz).

Plethysmography measurements represented the change in blood volume, so this signal could be used to estimate beat-to-beat intervals. Heart rate variability (HRV) signal was calculated using the `hrvanalysis` [19] Python library from beat-to-beat intervals extracted from the plethysmography signal.

TABLE 1: EXTRACTED FEATURES, PSD – POWER SPECTRAL DENSITY, STD– STANDARD DEVIATION, VLF – VERY LOW FREQUENCY, LF – LOW FREQUENCY, HF – HIGH FREQUENCY, RMS – ROOT MEAN SQUARE, SCR – SKIN CONDUCTANCE RESPONSE

Sign.	Features	Total = 1490
EEG [20]	For raw and normalized (range [0-1]) EEG signals: **Mean, Standard deviation, Mean of first derivative, Mean of second derivative**	32 ch x 8 features
	Hjorth features (Activity, Mobility, Complexity)	32 ch x 3 features
	For alpha, beta and theta band: **PSD** For raw and normalized (range [0-1]) EEG signals in alpha, beta and theta band: **Mean, Standard deviation, Mean of first derivative, Mean of second derivative**	32 ch x 3 bands x 9 features
	For alpha, beta and gamma band: **Energy, Recursive energy efficiency**	32 ch x 3 bands x 2 features
HRV [12,24]	**First derivative, Mean arc-length, RMS, Area-perimeter ratio, Mean and standard deviation, PSD in LF band - [0.01, 0.08] Hz, PSD in medium band - [0.08, 0.15] Hz PSD in HF band - [0.15, 0.5] Hz, PSD ratio between [0.04, 0.15] Hz and HF band**	10 features
Respiration [20]	**Maximum amplitude in frequency spectrum, Mean spectrum in [0.2, 0.5] Hz, Maximum amplitude in PSD, Mean PSD in [0.2, 0.5] Hz**	4 features
	Mean, Standard deviation, Median and Range of peak-to-peak intervals, STD of first derivative of intervals, Mean of breathing rate, STD of breathing rate, Maximum and minimum of breathing rate, Number of intervals larger than 50 and 20 ms and their ratio in total number of intervals, Square root of mean of sum of peak-to-peak intervals, Coefficient of interval change and variation, Total PSD, PSD in VLF range [0.003, 0.04] Hz PSD in LF) range [0.04, 0.15] Hz PSD in HF range [0.15, 40] Hz, LF/HF ratio, Normalized power in LF and HF domain	23 features
GSR [12, 20]	4 features for raw SCR and LF SCR: **Mean value, Standard deviation, Mean of first derivative, Mean of second derivative** 4 features for LF and VLF SCR: **Numbers of peaks in LF SCR, Number of peaks in VLF SCR, Number of peaks ratio in LF and VLF SCR, Mean of amplitude of LF and VLF SCR**	12 features
	Zero-crossing rate for LF SCR and VLF SCR	2 features
EMG [12, 20]	**PSD in [4, 40] Hz for zEMG and tEMG signals**	2 features
	4 features for raw and LF tEMG and 4 features for raw and LF zEMG signals: **Mean value, Standard deviation, mean value of first derivative, Mean value of second derivative** 3 features for LF and VLF tEMG: **Number of peaks in LF tEMG, Number of peaks in VLF tEMG, Number of peaks ratio in LF and VLF tEMG** 3 features for LF and VLF zEMG: **Number of peaks in LF zEMG, Number of peaks in VLF zEMG, Number of peaks ratio in LF and VLF zEMG**	22 features
Temp. [12]	**Mean, Standard deviation, First derivative, Minimum value, Maximum value, PSD in [0, 0.1] Hz, PSD in [0.1, 0.2] Hz**	7 features

C. Class labeling

Subjects gave ratings along multiple parameters after every video watched. Of these, we picked four parameters - valence, arousal, dominance and liking. Values for each parameter was in range [1, 9]. The classification problem was considered as the binary classification for each of the four previously mentioned parameters. Classes were defined as "0" and "1" which represented low and high parameter value, respectively. In this study, the class distinction boundary was set to 4.5. Classes were slightly imbalanced (around 60% of values fall in the high range).

We performed an analysis of the effect of changing the threshold on classification accuracy. Fig. 2 shows the results of training a Support Vector Machine (SVM) for each cut-off. As higher accuracy is achieved only on thresholds corresponding to an even higher-class imbalance (suggesting it is a result of the classifier always predicting the more common class), 4.5 was chosen as the boundary in the rest of our analysis.

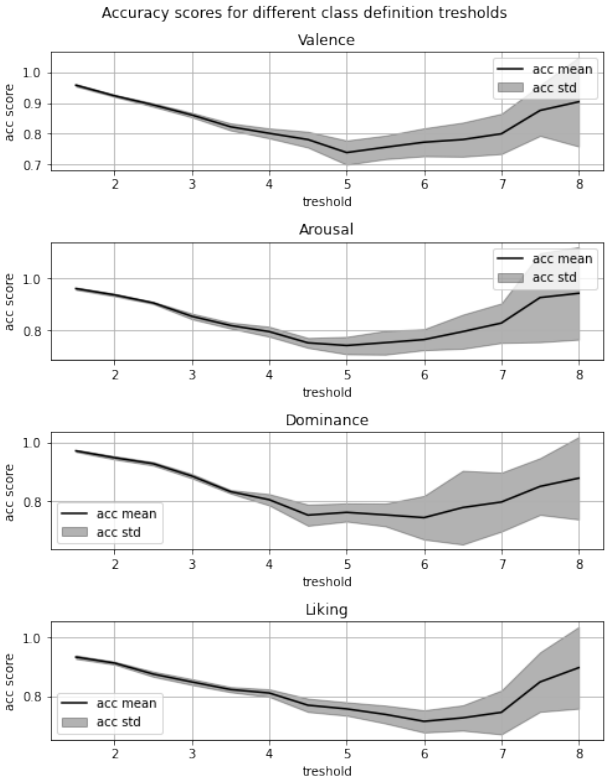

Fig. 2. Accuracy scores for different class definition thresholds.

D. Feature informativeness and dimensionality reduction

The importance for each feature and each emotion parameter was analyzed by mutual information between the feature and target variable, using the implementation available in the `sklearn` package [21].

As the final extracted set of features was of too high dimensionality for the size of available data: 1490 features vs 1280 recordings (32 subjects x 40 videos = 1280), dimension reduction was performed. The first step was removing repetitive features (when absolute correlation between two features is over 0.95, the feature which has a lower mutual information score with the target is removed).

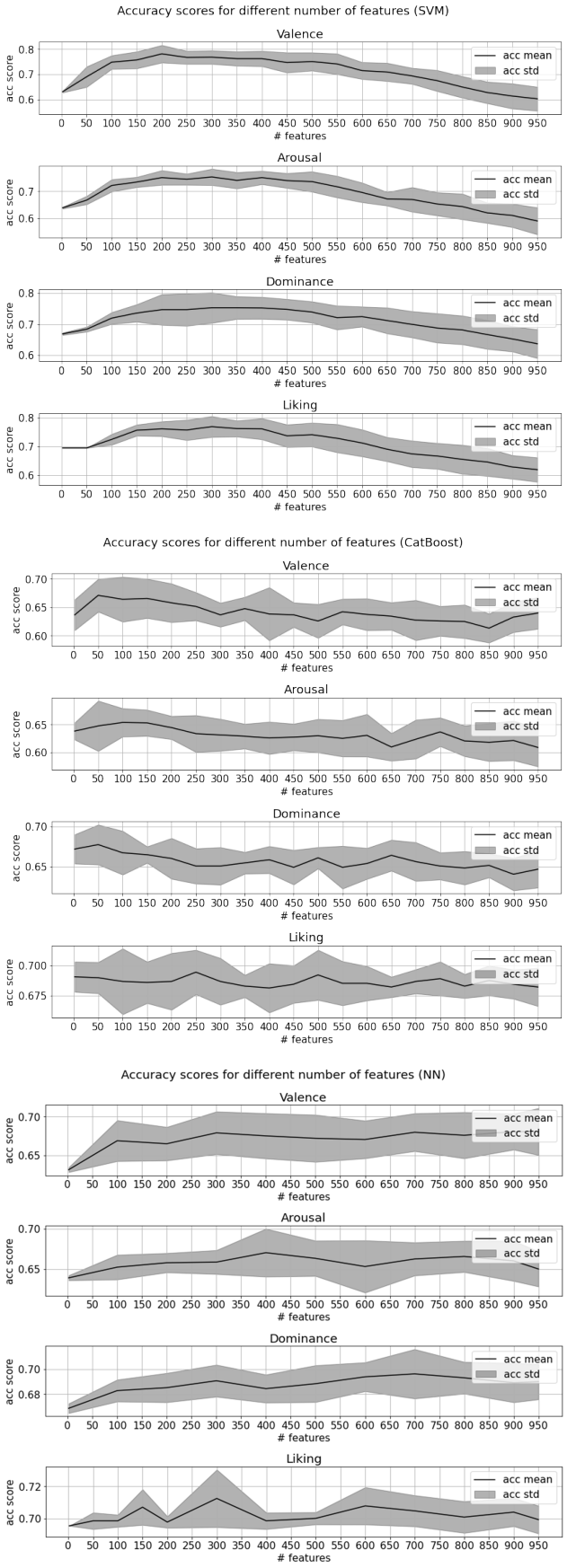

Fig. 3. Accuracy scores by number of features used SVM (top), CatBoost (middle) and ANN (bottom).

Next step was analyzing changes in classification accuracy when increasing the number of used features. This task was done for the SVM, CatBoost [22] and Artificial Neural Network (ANN) classifier. Fig. 3 shows mean cross-validation accuracy values and standard deviations achieved using SVM, CatBoost and ANN. The final set of features is decided based on maximum accuracy values for each emotion parameter and each classifier individually. The selection of features in each step for the first two classifiers was performed using the recursive feature elimination (RFE) algorithm. For the ANN, feature selection was performed by iteratively adding features ranked by their mutual information score. The network architecture in each iteration was adjusted so the size of the hidden layers is ¼ (1/2 for liking parameter) and 1/6 of the number of input features.

The final number of features for each machine learning algorithm is as follows:

- Catboost – 50 features for valence, 100 features for arousal, 50 features for dominance and 250 features for liking
- SVM – 200 features for valence and 300 features for arousal, dominance, and liking
- ANN – 400 features for arousal, 900 features for valence, 700 features for dominance, 300 features for liking.

E. Classification

After feature extraction and dimensionality reduction, the evaluation of different machine learning algorithms was performed.

The first evaluated method was Support Vector Machine which works by translating the feature space to a higher dimensionality one, where the data becomes linearly separable. The SVM implementation was realized using the `sklearn` package. Linear kernel was selected with 0.01 regularization parameter.

Boosting algorithms were considered for their historically good performance on tabular datasets. For this problem, `CatBoost` [25] Python package was chosen. Classifier parameters were selected using grid search – loss function was LogLoss with a learning rate of 0.1, a maximum tree depth of 3 and subsampling ratio of 0.8.

The third approach that was performed was ANN: a network with two hidden layers and a LeakyReLU activation function. Training was done for 50 epochs and the model that achieved the best validation accuracy was selected. Learning rate was set at 0.02, with a reduction by a factor of 0.5 every 5 epochs. Overfitting was resolved by using batch normalization and dropout regularization. The `PyTorch` [23] library was used for ANN implementation.

III. Results and Discussion

A. Correlation between emotion parameters

The correlation between each emotion parameter for all participants was calculated, along with the correlation between parameters and the videos (Video_id) participants were watching. The absolute values are shown in Fig. 4.

The most correlated parameters are valence and liking, and valence and dominance. Correlation with Video_id shows us how similarly the participants rated the videos for each emotion parameter. Arousal is uncorrelated to the Video_id, indicating there was great disagreement between participants when rating videos in this parameter.

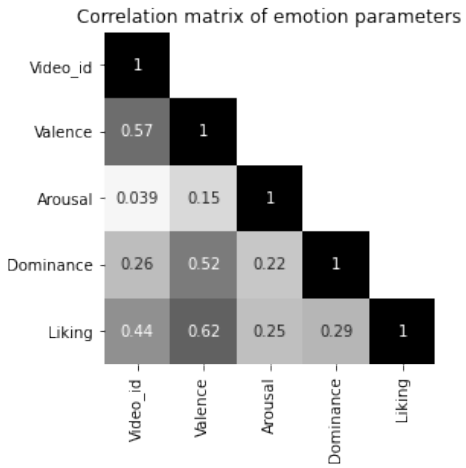

Fig. 4. Correlation matrix of emotion parameters.

B. The most informative features

Fig. 5 shows the top 10 features for each emotional parameter ranked by mutual information scores.

For arousal, the most important feature is the number of peaks in low-passed zEMG signal. Most of the important features come from EEG in the beta, theta, gamma, and alpha frequency band.

For valence, all of the important features are extracted from EEG signal. The most important feature is the mean value of the second derivative of signal from the F7 electrode in the alpha band.

For liking, most features come from EEG, combined with GSR and respiration.

For dominance, only one feature in the top 10 comes from EEG. The most important is the number of peaks in a low-passed tEMG signal. Other important modalities are zEMG and GSR.

C. Classification evaluation

Table 2 shows accuracy and F1 scores reached using SVM, CatBoost and ANN methods for each emotional parameter. All shown accuracies and F1 scores are mean values calculated on 10-fold stratified cross-validation (90:10 train-test ratio).

CatBoost and ANN reached very similar scores, while SVM performed the best for every parameter. The best accuracy was achieved for the valence parameter (78.1%), and the best F1 score for liking (85%). For arousal and dominance, the accuracies were somewhat lower (75.2% and 75.3% respectively). The full comparison is given in Table 2.

TABLE 2: CLASSIFIER EVALUATION

	Arousal (%)		**Valence (%)**		**Dominance (%)**		**Liking (%)**	
	Acc	***F1***	***Acc***	***F1***	***Acc***	***F1***	***Acc***	***F1***
SVM	**75.2**	**82.1**	**78.1**	**83.7**	**75.3**	**82.9**	**77.0**	**85.0**
Boost	65.4	77.0	67.1	77.6	67.7	79.8	69.5	81.7
ANN	67.0	77.9	68.0	77.8	69.6	80.4	71.3	82.5

These results outperform the classification described in Koelstra et al. [12] on the same dataset. Valence classification accuracy outperforms the one given by [11] (78.1% vs. 58%) while for arousal it is very similar (75.2% vs 75%).

IV. CONCLUSION

In this paper, we have compared the results of binary classification (high vs low) using different machine learning approaches (SVM, CatBoost, ANN) in case of four typical emotional parameters (arousal, valence, dominance and liking) on the publicly available DEAP dataset. An extensive set of features (1490) was generated from the available physiological signals. The final achieved classification accuracy was higher than previously reported, particularly when using the SVM classifier, which showed superior results in comparison to other two classifiers. The inability to further improve the accuracy might be due to the general nature of cross-subject emotion classification or an inherent problem within the reliability of data labeling according to subjective criteria.

Fig. 5. Mutual information for top 10 features.

Analysis of the ratings given by each subject shows great discrepancies in how the videos used in the experiment are perceived. This might be remedied by conducting an experiment choosing a different set of videos, particularly ones that have low rating deviations.

Feature analysis has shown that EEG signals carry the most information, but other modalities are not to be discarded, as for the dominance parameter electromyography and GSR carry the most information.

An important limitation of this analysis is the small number of subjects, influenced by the complexity and nonconformity of the experiment. Further research improvements could be made by applying the hybrid approach to the analysis, based on the combination of physiological and non-physiological (face expression) data.

REFERENCES

[1] D. B. Lindsley, "Emotion," *Handbook of experimental psychology,* pp. 473-516, 1951.

[2] M. Li, H. Xu, X. Liu and S. Lu, "Emotion recognition from multichannel EEG signals using K-nearest neighbor classification," *Technology and Health Care,* vol. 26, pp. 509-519, 2018.

[3] Y. L. Lin and G. Wei, "Speech emotion recognition based on HMM and SVM," in *2005 International Conference on Machine Learning and Cybernetics*, Guangzhou, China, 2005.

[4] Z. Liu, M. Wu, W. Cao, L. Chen, J. Xu, R. Zhang, M. Zhou and J. Mao, "A facial expression emotion recognition based human-robot interaction system," *IEEE/CAA Journal of Automatica Sinica,* vol. 4, pp. 668-676, 2017.

[5] S. Jerritta, M. Murugappan, K. Wan and S. Yaacob, "Emotion recognition from facial EMG signals using higher order statistics and principal component analysis," *Journal of the Chinese Institute of Engineers,* vol. 37, pp. 385-394, 2014.

[6] Y. L. Hsu, J. S. Wang, W. C. Chiang and C. H. Hung, "Automatic ECG-based emotion recognition in music listening," *IEEE Transactions on Affective Computing,* vol. 11, pp. 85-99, 2017.

[7] C. Lee, S. Yoo, Y. Park, N. Kim, K. Jeong and B. Lee, "Using neural network to recognize human emotions from heart rate variability and skin resistance," in *IEEE Engineering in Medicine and Biology 27th Annual Conference*, Shanghai, China, 2005.

[8] Y. Huang, J. Yang, P. Liao and J. Pan, "Fusion of facial expressions and EEG for multimodal emotion recognition," *Computational intelligence and neuroscience,* pp. 1-8, 2017.

[9] Y. Tan, Z. Sun, F. Duan, J. Solé-Casals and C. F. Caiafa, "A multimodal emotion recognition method based on facial expressions and electroencephalography," *Biomedical Signal Processing and Control,* vol. 70, 2021.

[10] R. Li, Y. Liang, X. Liu, B. Wang, W. Huang, Z. Cai, Y. Ye, L. Qiu and J. Pan, "MindLink-Eumpy: An Open-Source Python Toolbox for Multimodal Emotion Recognition," *Frontiers in human neuroscience,* vol. 15, 2021.

[11] C. A. Torres-Valencia, H. F. García-Arias, M. A. Álvarez López and A. A. Orozco-Gutiérrez, "Comparative analysis of physiological signals and electroencephalogram (EEG) for multimodal emotion recognition using generative models," in *2014 XIX Symposium on Image, Signal Processing and Artificial Vision*, Armenia, Colombia, 2014.

[12] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing,* vol. 3, pp. 18-31, 2011.

[13] H. Yang, J. Han and K. Min, "A Multi-Column CNN Model for Emotion Recognition from EEG Signals," *Sensors,* vol. 19, 2019.

[14] T. Stajić, J. Jovanović, N. Jovanović and M. Janković, "Emotion Recognition Based on DEAP Database Physiological Signals," in *2021 29th Telecommunications Forum (TELFOR)*, Belgrade, Serbia, 2021.

[15] C. Harris et al., "Array programming with NumPy," *Nature,* p. 357–362, 2020.

[16] P. Virtanen et al., "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature methods,* vol. 17, pp. 261-272, 2020.

[17] A. Bizzego, A. Battisti, G. Gabrieli, G. Esposito and C. Furlanello, "Pyphysio: A physiological signal processing library for data science approaches in physiology," *SoftwareX,* vol. 10, 2019.

[18] B. Hjorth, "EEG analysis based on time domain properties," *Electroencepalography and Clinical Neurophysiology,* vol. 29, pp. 306-310, 1970.

[19] R. Champseix, "Aura-healthcare/hrv-analysis: Package for Heart Rate Variability analysis in Python," Association AURA, [Online]. Available: https://github.com/Aura-healthcare/hrv-analysis. [Accessed 14 July 2021].

[20] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening",," *IEEE transactions on pattern analysis and machine intelligence,* vol. 30, pp. 2067-2083, 2008.

[21] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *The Journal of Machine Learning Research,* vol. 12, pp. 2825-2830, 2011.

[22] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Advances in neural information processing systems 31 (NeurIPS 2018)*, Montréal, Canada, 2018.

[23] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Vancouver, Canada, 2019.

[24] A. Bartolomé-Tomás, R. Sánchez-Reolid, A. Fernández-Sotos, J. M. Latorre and A. Fernández-Caballero, "Arousal Detection in Elderly People from Electrodermal Activity Using Musical Stimuli," *Sensors,* vol. 20, 2020.