Prevent Over-fitting and Redundancy in Physiological Signal Analyses for Stress Detection

Ruijie Fang
ECE Department
University of California, Davis
CA, USA
rjfang@ucdavis.edu

Ruoyu Zhang
ECE Department
University of California, Davis
CA, USA

Elahe Hosseini
ECE Department
University of California, Davis
CA, USA

Anna M. Parenteau
Department of Psychology
University of California, Davis
CA, USA

Sally Hang
Department of Psychology
University of California, Davis
CA, USA

Setareh Rafatirad
ECE Department
University of California, Davis
CA, USA

Camelia E. Hostinar
Department of Psychology
University of California, Davis
CA, USA

Mahdi Orooji
ECE Department
University of California, Davis
CA, USA

Houman Homayoun
ECE Department
University of California, Davis
CA, USA

2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9995121

***Abstract*—Stress detection is an emerging field. WESAD is a commonly used public dataset for automated stress detection. It contains physiological signals including ECG, EDA, EMG, ACC, BVP, EDA, and skin temperature. The time window approach is used to extract features from time-series physiological signals. We find in previous studies that a 60-second time window with a 0.25-second window shift is widely used, but such window settings may cause redundancy and over-fitting. Thus, we propose to use (1) new window settings and (2) normalization per subject to tackle this problem. The experiment results show that our proposed methods significantly increase the classification performance.**

***Index Terms*—Automated Stress Detection, Machine Learning, Multi-modal Learning, Affective Computing, Emotion Recognition**

I. Introduction

Affective computing has attracted increased interest in recent decades as it may offer empathy-like ability to machines [1]–[6], which is important for human-computer interaction [7], [8]. Stress, defined as the "nonspecific response of the body to any demand upon it" [9], is an important target for disease prevention. Stress can cause depression, insomnia, and gastrointestinal problems. The stress response involves the activity of two primary systems: the sympathetic–adrenal–medullary (SAM) system and the hypothalamic–pituitary–adrenal (HPA) axis [10]. The SAM system works as follows: when an individual is feeling stressed, the brain sends a message to the autonomic nervous system (ANS), which triggers the adrenal glands to release noradrenaline, followed by a higher heart rate, quicker breathing, and other rapid and intense changes. This is referred to as the "fight-or-flight" response [11].

This stress mechanism allows engineers to develop automated stress detection algorithms by detecting physiological changes that correspond to SAM responses. Numerous datasets have been collected and published for stress detection. They focus on video and physiological signals representing an individual's SAM-related activity. WESAD is a well-known dataset that collected 15 subjects' data. The experiment procedure includes twenty minutes baseline, funny video watching for around six minutes, ten minutes TSST, meditation, and recovery. The procedure yields five tags baseline, amusement, meditation, stress (induced via Trier Social Stress Test), and rest. During the entire experiment process, physiological signals were collected, including electrocardiogram (ECG), electrodermal activity (EDA), blood volume pulse (BVP), electromyogram (EMG), and respiration, body temperature, and three-axis acceleration. The baseline and stress are two attractive classification targets in the automated stress detection field; previous research attempted to solve this binary classification problem by applying machine learning approaches.

When machine learning techniques are deployed to detect stress using WESAD, the time window is the typical method for feature extraction, especially for traditional machine learning. There are two types of time windows: overlapping time windows and non-overlapping time windows. As the name suggested, an overlapping time window has overlapping parts between consecutive windows, as shown in Figure 1 and non-overlapping windows are split end to end as shown in Figure 2. Schmidt et al. [1] first introduced the WESAD dataset and applied 60 second time window and 0.25-second shift when introducing the WESAD dataset. Gil-Martin et al. [12] also used 60-second windows with 0.25-second window shift and applied convolutional neural networks. 96.6% accuracy was obtained in detecting stress v.s. non-stress. Heo et al. [13] deployed fancy PPG processing and feature extraction techniques and attained 0.77 area under the curve (AUC) of the receiver operating characteristic (ROC) curve by only using PPG for 3-class classification. A time window size of 2 minutes and a window shift of 0.25 seconds were used. However, with the over-lapping window settings illustrated above, we suspect information is redundant and overfitting is happening due to high-level overlapping (over 99%). Thus, we conducted a series of experiments to verify our concern and proposed a framework consisting of adaptive overlapping windows and normalization per subject to solve the proposed problem.

Our main contributions can be summarized as:

- We verify our hypothesis that a largely overlapping time window causes over-fitting in machine learning models and is infeasible for application.
- We propose two methods to solve this issue. First, we apply adaptive overlapping windows and find that specifically, the 30s, which is 50% overlapping was the best and then, we deploy normalization data per subject to further weakend the data bias.
- Experiments are conducted to verify the effectiveness of proposed methods.

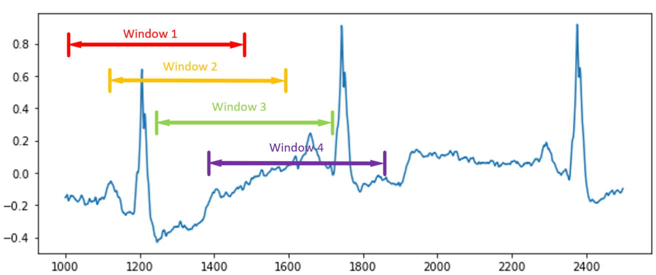


Fig. 1. Overlapping time window

Fig. 2. Non-overlapping time window

The rest of this paper is organized as follows: Section II explains the details of the proposed methods. Then, experiments are conducted and results are presented in Section III. In Section IV we discuss the advantages and weakness of this study and propose future research directions. Finally, conclusions are discussed in Section V.

II. METHODS

In this section, we propose two approaches to fix the over-fitting issue including lower overlapping windows and normalization per subject.

A. Leave Subjects Out

We predict that higher overlapping time windows are causing over-fitting due to consecutive samples going to training or testing sets. We tested our hypothesis by using the ”leave subjects out” (LSO) experiment. The proposed approach works as follows: when splitting training and testing sets, instead of randomly splitting the entire mixed dataset, we leave around 30% subjects as the testing set and the rest 70% subjects as the training set. Through this method, we eliminate the overlap between the training set and testing set because only windows from the same subject can overlap, and different subjects’ data have no chance of overlapping. We conducted experiments based on it, and are explained in Section III.

B. Adaptive Overlapping Windows

Intuitively, if higher overlapping leads to over-fitting, lower overlapping is worth trying as less similar information is going to be distributed to the training set and testing set. Here, we propose to use mutual information to adapt time window by finding the specific time window with highest mutual information between features and labels. Mutual information is defined as:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

C. Normalization per Subject

In machine learning, normalization is useful especially when Euclidean distance is involved, e.g., support vector machine. The normalization equation is illustrated in equation 2.

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{2}$$

It is normally done on the entire mixed dataset, which contains data from all subjects. However, in the field of affective computing, individuals are different in terms of their physiological baseline and stress responses. Thus, traditional normalization cannot help information extraction. We propose to normalize data per subject so that the dissimilarity can be initially reduced. Fig. 3 and Fig. 4 shows that how two normalization works and the differences between traditional normalization and normalization per subject. We also conducted experiments to test the effectiveness of this approach which will be further discussed in Section III.

III. EXPERIMENT AND RESULTS

In this section, we present the experiments we conducted and results.

In all experiments, we used three physiological signals collected in the WESAD dataset: electrocardiogram (ECG), electrodermal activity (EDA), and electromyogram (EMG). All signals were sampled at 700 Hz, and we used time window methods to extract features. All features come from Schmidt et al.’s work, and for detailed feature information, please refer to [1] as feature extraction is not emphasized in this study. In total, 14 features from ECG, 14 from EDA, and 14

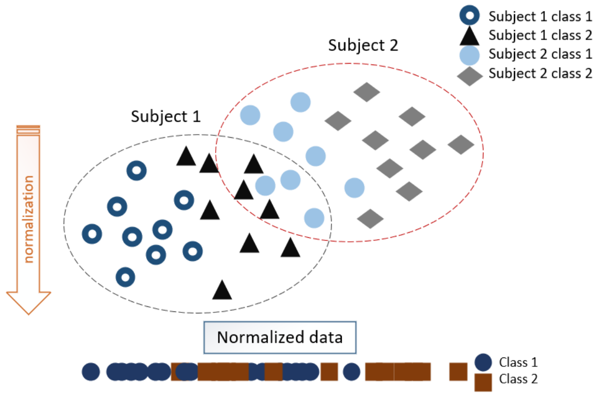


Fig. 3. Traditional normalization

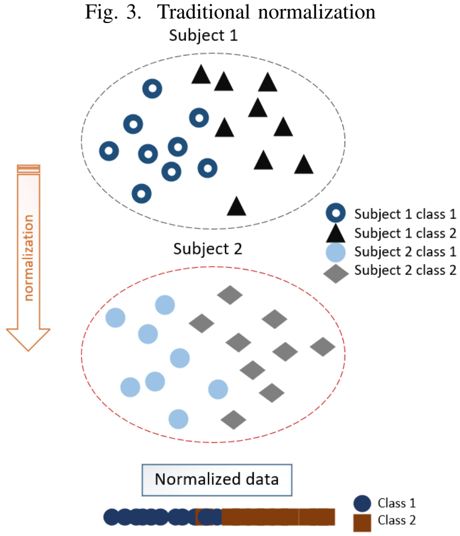


Fig. 4. Normalization per subject

from EMG were extracted and used throughout the study. In addition, the classification algorithms used are random forest, support vector machine with "RBF" kernel, and artificial neural network. The hyperparameters of models are fine-tuned by using the GridSearchCV algorithm. The evaluation metrics are the median of three algorithms.

TABLE I
CLASSIFICATION RESULTS

Experiment	Accuracy	F-1 Score	AUC
60/0.25 LSO	69.0%	62.3%	0.739
60/30 LSO	78.8%	77.9%	0.914
NPS LSO	97.8%	96.9%	0.998

First of all, We applied a mutual information computation with a time window length from 10 to 120s with a 10-second step and time window shift from 0.25s minimum to time window length maximum with a 0.25s step and found that 60 seconds time window with 30 second time window shift had the highest mutual information. We conducted the same experiments for both 60s/0.25s settings and 60s/30s to verify our assumption that too overlapping causes redundancy and overfitting. Results are presented in Table I and Figure 5 (b). We found that using lower overlapping windows can weaken the over-fitting effects and improve the performance. By doing so, the classifier accuracy has been increased from 69% to 78.8%.

Then, we deployed the normalization per subject approach that normalized data for each subject before they were mixed into the dataset. A sample data change shows the significant influence of normalization per subject and is shown in Fig. 6 and Fig.7. All data samples shown on the figure come from two subjects' data in the WESAD dataset. Two features, EDA mean absolute value and HRV mean NN interval are used as x and y axis. Before normalization per subject, subject A's label 1 data (red) and subject B's label 0 data (blue) are entangled. After normalization per subject, there can be a straight line drawn between green, blue (label 1 from AB) and red, orange (label 0 from AB) which means the data are more discriminative than before. Promising results can be found in Table I and ROC curve in Figure 5 (c). The accuracy increased from 78.8% to 97.8% which is consistent with our hypothesis and with the 2-d sample feature space figure.

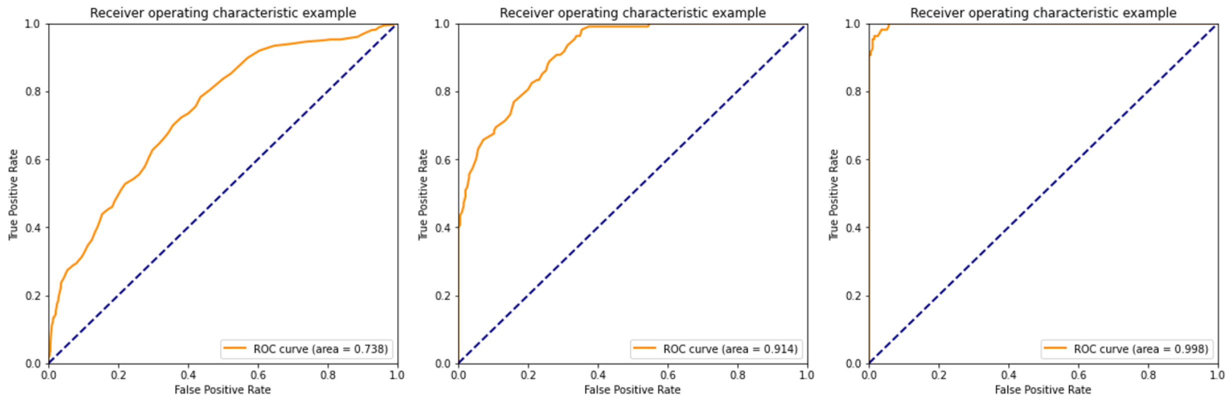
IV. DISCUSSION

In the adaptive overlapping method we proposed, the classification performance significantly increased, which illustrated that adaptive overlapping, specifically lower overlapping, can prevent over-fitting and information redundancy. However, there is a trade-off between higher overlapping and lower overlapping: with higher overlapping, there can be more data samples generated, e.g., with a 0.25s time window, over 66,000 data samples can be used, while with a 30s time window, only 550 data samples. It is clear that with more data samples, the classifier gets more robust because of the regularization of data. However, more overlapping data means more data reuse and redundancy. Thus, it is a trade-off and needs to be considered every time.

In terms of normalization, our results showed that it is significantly helpful if we normalize per subject instead of normalizing the entire dataset. This is not a specific method for WESAD or physiological signals, but it can be extended to other datasets and other modalities, e.g., audio and video, as the principle of such method is general, as shown in Fig. 3 and Fig. 4, and as shown in Fig. 6 and Fig. 7, before normalization, data are mixed in the EDA MAV-HRV MeanNN space, and it is not clearly separated. After normalization per subject, label 0 and label 1 data can be easily discriminated by eyes.

V. CONCLUSION

In this paper, we tested our hypothesis that the commonly used high overlapping time window approach can lead to overfitting when training machine learning models. We conducted experiments, and our results support our hypothesis. Then, we proposed two methods to prevent over-fitting: (1) lower overlapping and (2) normalization per subject. The two proposed methods were verified to be effective by conducting



(a) 0.25s time window (b) 30s time window (c) normalization per subject

Fig. 5. ROC from experiments

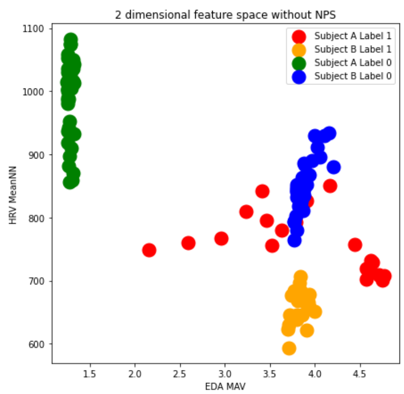


Fig. 6. 2-dimensional sample feature space in WESAD without normalization

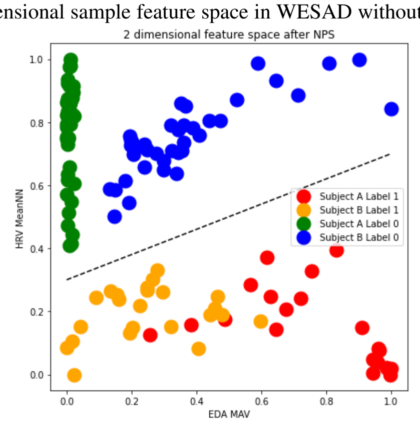


Fig. 7. 2-dimensional sample feature space in WESAD after normalization per subject

experiments, and the classification accuracy was upgraded from initially 69.0% to 78.8% and, in the end, increased to 97.8%.

References

[1] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 400–408, 2018.

[2] Nandita Sharma and Tom Gedeon. Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Computer methods and programs in biomedicine*, 108(3):1287–1301, 2012.

[3] Ruijie Fang, Ruoyu Zhang, Sayed M Hosseini, Mahya Faghih, Soheil Rafatirad, Setareh Rafatirad, and Houman Homayoun. Pain level modeling of intensive care unit patients with machine learning methods: An effective congeneric clustering-based approach. In *2022 4th International Conference on Intelligent Medicine and Image Processing*, pages 89–95, 2022.

[4] Arshad Hashmi and SK Yadav. A systematic review of computational methods for occupational stress modeling based on subjective and objective measures. *Int J Comp Sci Eng 2018; 6: 456*, 70, 2018.

[5] M Arif, A Basri, G Melibari, T Sindi, N Alghamdi, N Altalhi, and M Arif. Classification of anxiety disorders using machine learning methods: a literature review. *Insights Biomed Res*, 4(1):95–110, 2020.

[6] Ruijie Fang, Ruoyu Zhang, Elahe Hosseini, Sayed Hosseini, Mahya Faghih, Mahdi Orooji, Soheil Rafatirad, Setared Rafatirad, and Houman Homayoun. Atlas: An adaptive transfer learning based pain assessment system: A real life unsupervised pain assessment solution. *44th International Engineering in Medicine and Biology Conference*, 2022.

[7] Stanisław Saganowski, Anna Dutkowiak, Adam Dziadek, Maciej Dzieżyc, Joanna Komoszyńska, Weronika Michalska, Adam Polak, Michał Ujma, and Przemysław Kazienko. Emotion recognition using wearables: A systematic literature review-work-in-progress. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 1–6. IEEE, 2020.

[8] Ayten Ozge Akmandor and Niraj K Jha. Keep the stress away with soda: Stress detection and alleviation system. *IEEE Transactions on Multi-Scale Computing Systems*, 3(4):269–282, 2017.

[9] Hans Selye. Stress without distress. In *Psychopathology of human adaptation*, pages 137–146. Springer, 1976.

[10] Daryl B O'Connor, Julian F Thayer, and Kavita Vedhara. Stress and health: A review of psychobiological processes. *Annual review of psychology*, 72:663–688, 2021.

[11] Margaret E Kemeny. The psychobiology of stress. *Current directions in psychological science*, 12(4):124–129, 2003.

[12] Manuel Gil-Martin, Ruben San-Segundo, Ana Mateos, and Javier Ferreiros-Lopez. Human stress detection with wearable sensors using convolutional neural networks. *IEEE Aerospace and Electronic Systems Magazine*, 37(1):60–70, 2022.

[13] Seongsil Heo, Sunyoung Kwon, and Jaekoo Lee. Stress detection with single ppg sensor by orchestrating multiple denoising and peak-detecting methods. *IEEE Access*, 9:47777–47785, 2021.