Driving stress detection using physiological data with machine learning

Bui Tien Dat, Tran Quang Duc, Cung Thanh Long*

Hanoi University of Science and Technology.
*Corresponding author: long.cungthanh@hust.edu.vn

Received 02 Jul 2022; Revised 06 Sep 2022; Accepted 07 Nov 2022; Published 18 Nov 2022.
DOI: https://doi.org/10.54939/1859-1043.j.mst.83.2022.22-29

ABSTRACT

Stress is a problem that affects both physical and mental health, causing negative emotional states. Stress can impair the driver's ability to perceive and handle situations in driving safety. Therefore, the detection and assessment of stress levels play an important role in improving comfort, well-being, and enhancing the driving experience for drivers. Using the AffectiveROAD dataset, this paper proposes a method of classifying stress levels through physiological signals obtained from driving sessions. These signals are time-aligned and pre-processed to extract the suitable features within a five-second period. Based on the obtained features, Machine Learning models are trained to classify stress status into five levels. The tested results show that the accuracy reaches 94% with the Random Forests (RF) when using the seven most important features from the HR, EDA, TEMP signals, and 99% when incorporating the overlapping technique for 10-fold cross-validation.

Keywords: Stress detection; Wearable sensors; Feature extraction; Machine learning; Random forests.

1. INTRODUCTION

Stress is the body's response to any request, pressure, or impact factor affecting a person both physically and mentally. Signs of stress include mood changes, decreased sex drive, sleeping difficulty, low energy, muscle tension, racing heartbeat, etc. Stress can have both positive and negative effects. On the positive side, stress helps the body stay awake and focused, as well as stimulates strong activities of the senses, such as the heart beating faster to supply blood to the brain and muscles. However, in most cases, stress brings negative effects and adversely affects human health, resulting in, for example, diabetes, hair loss, obesity, heart disease, hyperthyroidism, and sexual dysfunction. These impacts can negatively influence daily life, including diminished work performance and distraction, both of which lead to an increased risk of occupational accidents and other health-related problems [1]. In recent years, the ratio of traffic accidents related to drivers' stress problems has increased remarkably [2]. Therefore, the task of detecting and assessing an individual's stress level makes a significant contribution to people in modern life and has many important applications such as monitoring, health care, ensuring safe driving [3–5]. However, implementation is unfeasible when stress is diagnosed through signs such as psychological signals, emotions, and behaviors. This information requires sophisticated sensors and facilities to gather and analyze.

Currently, only physiological experts and doctors can diagnose stress levels using random questions or tests. The disadvantages of this method are low accuracy and high dependence on subjective opinions. However, this task has become faster and easier thanks to the application of Machine Learning and Deep Learning algorithms. For instance, to detect stress for students under stress due to university examinations, Paolo Melillo [6] proposed a classifier for automatic stress detection based on nonlinear features of Heart Rate Variability (HRV). The author collected data from 42 students over two time periods (one during an ongoing university examination, assumed as a real-life stressor, and one after the holidays). The HRV signals are then extracted into 13 nonlinear features within a 5-minute period. Two labels will classify the stress level of the students as "non-stress" and "stress" using the Linear Discriminant Analysis (LDA) model, which enables stress detection with a total classification accuracy, sensitivity, and a specificity rate of 90%, 86% and 95%, respectively. Unlike the above model, which uses only the HRV signal, the classification model of Jing Zhai and Armando Barreto [7] uses information from four signals, namely galvanic skin response (GSR), blood volume pulse (BVP), pupil diameter (PD) and skin temperature (ST). By using eleven extracted features and a Support Vector Machine model, a consistent classification accuracy of 90% was achieved. The results also concluded that the PD signal had the greatest effect on stress states. Moreover, Feng-Tso Sun [8] proposed classifying stress according to different states of the subject, including standing, sitting and walking. The dataset includes signals such as GSR, electrocardiogram (ECG), and ACC data gathered from 20 participants. Classification accuracy reaches 92.4% in mental stress for 10-fold cross-validation, and 80.9% for between-subject classification when using activity information derived from the accelerometer. Alberto de Santos Sierra [9] introduced a system based on fuzzy logic that provides information on the mental state of an individual in short periods based on the characteristics of HR and GSR signals. The stress-detection accuracy reaches 99.5% for the 10-second period classification, and 90% for the 3-to-5-second period.

Pekka Siirtola [10] compared regression and classification models for stress detection. The dataset used is AffectiveROAD, which consists of data obtained from the Empatica E4 sensors and continuous target variables. These signals are extracted into 119 features with a 60-second sliding window. The author has surveyed user-independent and personal stress detection. The results show that the performance of the regression model outperforms the classification model in the problem of "stress" or "non-stressed" classification. The two models' best user-independent results are 82.3% and 74.1%. Moreover, the results obtained from the personal model show that the relationship between biological signals and stress status varies not only between the study subjects but also between the sessions gathered from the same person.

In general, very few studies have examined changes in stress over short periods of time. In most of the published studies, authors investigate stress classification with a period longer than one minute. Furthermore, these methods classify only two statuses: "stress" and "non-stress".

In this paper, we propose an approach to the classification of five stress levels. The dataset is similar to [10], including signals such as Temperature (TEMP), Heart rate (HR), Electrodermal Activity (EDA) and Accelerometer (ACC). We implement a procedure to extract suitable features in 5-second intervals. Then, the cross-validation technique is applied for training/ validating/ testing machine learning models.

The following sections of this paper are organized as follows: Part II introduces the dataset, preprocessing, and feature extraction; Part III implements multi-class classification models using a 5-second sliding window and optimizes the subset of features with the most significant effect on stress; Part IV presents some conclusions and our future works.

2. PREPROCESSING AND FEATURE EXTRACTION

2.1. Experimental Dataset

AffectiveROAD dataset consists of data collected from nine drivers, whose names are NM, RY, BK, MT, EK, KGS, AD, GM, and SJ. However, GM's data was collected twice, and RY's and NM's were done three times. Therefore, the dataset contains data from 13 data-gathering sessions. Each participant wore three devices to collect biological signals. The two Empatica E4 were worn on the left and the right wrist. Zephyr Bioharness 3.0 chest belt was worn on the driver's chest. The data gathered by Bioharness contains HR, breathing rate (BR), posture, and activity. The E4 measures TEMP, HR, EDA, Blood volume pulse (BVP), ACC, and the time between individuals' heartbeats (IBI). The same as in [10], only Empatica E4 data from the right wrist is chosen to use in this paper.

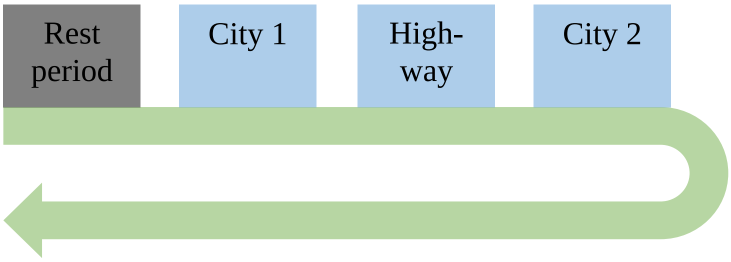

Figure 1. *Description of the Route.*

Data is collected on various road sections in normal traffic conditions, including rest periods (baseline), city driving and driving on the highway (figure 1). The stress level is assessed by continuous values between 0 (no stressful) and 1 (extremely stressful). Because there are no stress estimates available from baseline, we assume that during the whole baseline session, the level of stress is 0.

2.2. Training target

In the entire driving process, the stress level will fluctuate. Theoretically, when drivers are in complex traffic situations, such as in the city, where there are many other vehicles, traffic signals, and pedestrians, they are easily stressed. By contrast, they tend to loosen on a highway or in smooth traffic. However, the actual experiment shows that the stress index varies randomly depending on the characteristics of each individual as well as the gathered session as shown in figure 2. The stress s index of AD1 during highway sessions is higher than NM1 during city driving. Furthermore, the stress index tends to vary. Therefore, using multiple levels to assess stress values helps to track stress in a more detailed way compared to two levels (no stress and stress). This is different from [10], in which data from baseline are labeled as 0 and data from driving as 1.

In our work, stress values are classified into five groups (or five training targets):

- Non-stress (NS): stress metric from 0 to 0.2;
- Low stress (LS): stress metric from 0.2 to 0.4;
- Medium stress (MS): stress metric from 0.4 to 0.6;

- High stress (HS): stress metric from 0.6 to 0.8;
- Very high stress (VHS): stress metric from 0.8 to 1.

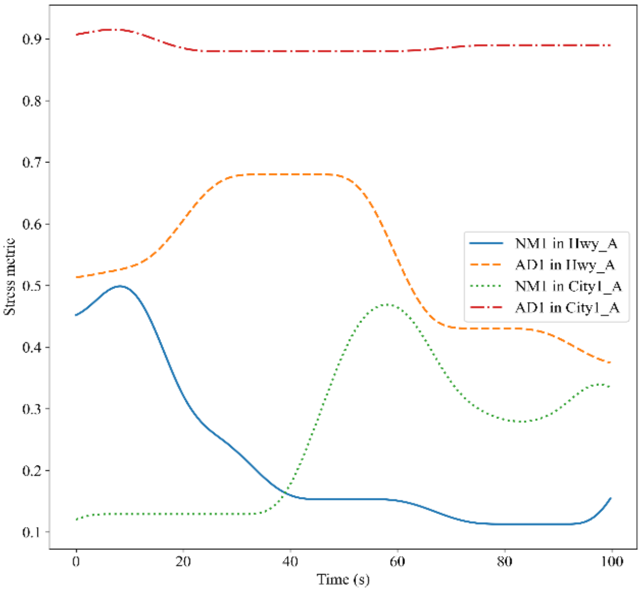

Figure 2. *Stress metric of AD1 and NM1 in City1 and Hwy session.*

The stress index often changes quickly in a short period, as in figure 2. Therefore, to ensure the output closely matches the signals used as model inputs, we divide the signals into small windows of the 5-second size. We also use the overlapping technique for data augmentation and reduce the gap between windows (with an overlapping rate of 80%). For a short window length, the stress values are guaranteed not to change so much in the whole window. The output value of the model will be the stress value at the end of the window.

2.3. Preprocessing

We concentrate on investigating the relationship between physiological signals on the right-hand Empatica E4 meter sensors and stress levels, including EDA, HR, TEMP, and ACC. The HRV signal is only meaningful when examined over a minimum period of a few minutes and is therefore not used [8]. Table 1 shows four used signals in our study.

Table 1. *Biologicals were used in this study.*

Signal	Information		
	Sampling Frequency (Hz)	Unit	Number of Features
HR	1	Hz	2
EDA	4	µS	11
TEMP	4	°C	6
ACC	32	$\frac{1}{64}g$	15

The raw signal is divided into several segments corresponding to the route, such as Rest, City1, City2, Hwy. They are divided into subsegments (windows) and then extracted from all desired features.

2.4. Feature Extraction

The HR, TEMP and ACC signals are extracted with the same features as in the paper [11]. The EDA signal is analyzed with the same statistical characteristics as the ST signal. In addition, the EDA signal is split into two signals: tonic (SCR) and phasic (SCL) signals. The SCR component exhibits a fast response to the stimulus, whereas the SCL component shows a stable and slowly variable composition [11]. The remaining features are the area under the identified SCRs, mean, std of SCR, and SCL. Details of EDA decomposition and algorithms to calculate features are available in the Neurokit 2 library [12]. The sum of all extracted features is 34. The features extracted from the different modalities are shown in table 2.

Table 2. *List of extracted features.*

Signal	Features	Description
HR	μ_{HR}, σ_{HR}	Mean, standard deviation of the HR
TEMP	$\mu_{TEMP}, \sigma_{TEMP}$, $\min_{TEMP}, \max_{TEMP}, \mathrm{range}_{TEMP}, \delta_{TEMP}$,	Mean, standard deviation of the TEMP, min, max TEMP, dynamic range, slope
ACC	$\mu_{ACC,i}, \sigma_{ACC,i}$, $i \in \{x, y, z, 3D\}$ $\left\| \int ACC, i \right\|, i \in \{x, y, z, 3D\}$ $f^{peak}_{ACC,i}, i \in \{x, y, z\}$	Mean, standard deviation for each axis separately and summed over all axes, Absolute integral for each/all axes Peak frequency for each axis i
EDA	μ_{EDA}, σ_{EDA}, $\min_{EDA}, \max_{EDA}, \mathrm{range}_{EDA}, \delta_{EDA}$, $\mu_{SCL}, \sigma_{SCL}, \mu_{SCR}, \sigma_{SCR}, SCR_{Area}$,	Mean, standard deviation of the EDA, min, max EDA, dynamic range, slope, Mean, standard deviation of the SCL/SCR, Area under the identified SCRs

3. IMPLEMENTATION, RESULTS AND DISCUSSIONS

In this work, we use multi-label classification models because they are suitable to the data and our training targets. Furthermore, we focused on investigating the relationship between biological signals and stress levels. The ACC signal is for reference and comparison with the original paper [10]. Since the ACC signal is specifically related to subject activity, this removal helps to generalize the application of the problem to more than just driving.

To find the optimal model, we survey machine learning models, including Random Forests (RF), Linear Discriminant Analysis (LDA), K-nearest neighbors algorithm (KNN), Quadratic Discriminant Analysis (QDA), Support Vector Machine (SVM), Multilayer Perceptron (MLP), etc (as listed in table 3). The data is divided into 5-second sliding windows without overlapping, and includes 34 features extracted from

HR, EDA, TEMP and ACC signals. Data important features selection and model formulation - training - testing are carried out in a Python environment with the help of “Sklearn”. 10-fold cross-validation was performed five times to evaluate the classification performance.

Table 3. *Accuracy of classification models.*

Model	Accuracy (%)							
	Number of features							
	34	30	25	20	15	10	7	5
RF	87.14	88.23	89.14	89.8	91.19	92.88	93.61	93.17
LDA	63.51	63.32	63.55	63.27	62.74	62.79	58.63	58.68
QDA	19.96	25.3	18.25	18.77	60.7	62.12	67.71	68.01
KNN	76.37	73.37	76.42	78.13	78.37	81.84	89.93	89.88
SVM	67.41	67.46	67.46	68.51	68.51	68.51	68.11	68.35
MLP	75.55	76.46	68.65	75.04	74.85	71.61	70.69	71.67
Bagging	77.22	77.17	78.7	78.8	78.13	77.37	86.01	80.21
Decision Tree	82.08	81.66	82.04	82.47	85.61	86.23	88.39	88.34
Extra Trees	84.4	83.84	86.2	87	88.61	88.23	59.61	61.51
AdaBoost	71.13	70.95	71.53	71.09	71.32	72.42	66.25	65.33
Gradient Boosting	78.08	76.66	78.04	78.13	78.08	76.99	88.39	77.33

Table 3 shows that the Random Forests model has the highest classification accuracy of 93.61% when using the seven most important features shown in figure 3. For these features, we do not use the ACC signal, which proves that our model does not need to use information from the ACC signal. The ACC signal received from the sensor on the wristwatch greatly depends on the activities of the drivers. For the most part, the driver's hand condition on the baseline and highway is usually stable, leading to a relatively small stress level. In contrast, on roads like City1 and City2, the driver must change the movement of his hands on the steering wheel constantly as he must focus on observing pedestrians, lights, and other traffic, which may relate to a high level of stress (but not always correct). So, detecting stress levels without using ACC signals also opens more real-world applications beyond just driving.

Because the number of training/testing data for each stress level is different and is shown in the confusion matrix in figure 4, we also use F1-score as an evaluation metric. F1-score is the harmonic mean of precision and recall and gives a better measure of the incorrectly classified cases than the accuracy metric, so it is recommended for unbalanced classification tasks.

As shown in figure 4, when applying an overlapping technique with a distance between predictions of 1 second, the RF model gives an improved accuracy result of up to 98.96% and an F1-score of 99% with the same optimized feature set (including seven most important features). The F1-score results are very high, demonstrating the good performance of the model.

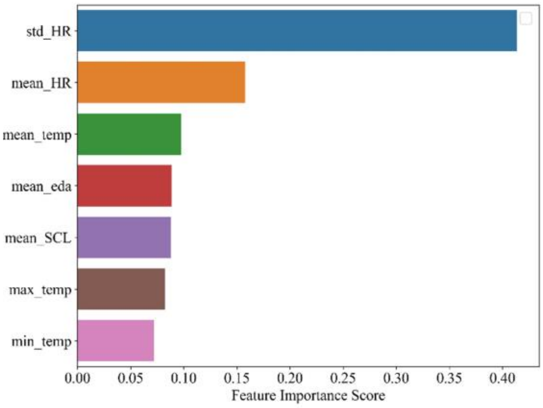

***Figure 3**. The importance of features.*

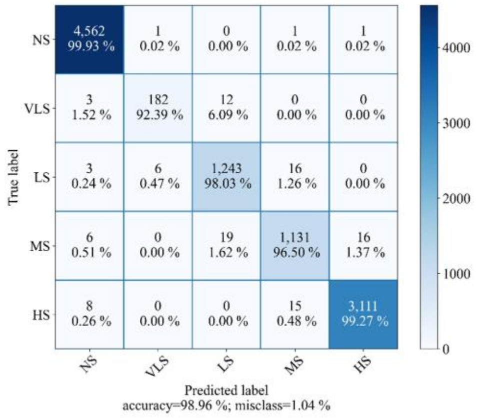

***Figure 4**. Confusion matrix with overlapping technique.*

4. CONCLUSION AND FUTURE WORKS

In this paper, we present an approach to assess the stress level of drivers using only the biological signals HR, TEMP and EDA without the ACC signal. This means that the model can be applied to many different contexts, such as stress prediction in school, work, etc. The paper also shows the different influences of signal features and shows our optimized set of features which gives high accuracy of classification results. This helps to reduce the complexity of the model and the amount of computation, making the model capable of embedding directly on wearable devices.

For our problem of five-level stress detection, we analyzed and worked on short (few-second) data segments. Tested results show a classification accuracy of 94% when using the optimal set of features and up to 99% when combined with the data overlapping technique.

In fact, the relationship between biological signals to stress status depends on the context and other conditions, so the coefficients of the classification model need to be updated and adapted quickly over time. This may require a continuous evaluation from the surveyor, and the data gathered increases continuously over time. In the next studies, we will focus on studying this kind of problem, using adaptive and semi-supervised algorithms for unlabeled data. In addition, the relationship between stress and other biological signals such as RES, EMG, will also be explored.

REFERENCES

[1]. L. Petitta, T. M. Probst, V. Ghezzi, and C. Barbaranelli, *"Economic stress, emotional contagion and safety outcomes: A cross-country study,"* Work, **vol. 66**, no. 2, pp. 421–435, (2020).

[2]. A. Kashevnik, I. Lashkov, and A. Gurtov, *"Methodology and Mobile Application for Driver Behavior Analysis and Accident Prevention,"* IEEE Trans. Intell. Transp. Syst., **vol. 21**, no. 6, pp. 2427–2436, (2020).

[3]. N. El Haouij, J.-M. Poggi, R. Ghozi, S. Sevestre-Ghalila, and M. Jaïdane, *"Random forest-based approach for physiological functional variable selection for driver's stress level classification,"* Stat. Methods Appl., **vol. 28**, no. 1, pp. 157–185, (2019).

[4]. M. Lohani, B. R. Payne, and D. L. Strayer, *"A Review of Psychophysiological Measures to Assess Cognitive States in Real-World Driving,"* Front. Hum. Neurosci., **vol. 13**, p. 57, (2019).

[5]. A. Němcová *et al.*, *"Multimodal Features for Detection of Driver Stress and Fatigue: Review,"* IEEE Trans. Intell. Transp. Syst., pp. 1–20, (2020).

[6]. P. Melillo, M. Bracale, and L. Pecchia, *"Nonlinear Heart Rate Variability features for real-life stress detection. Case study: students under stress due to university examination,"* Biomed. Eng. Online, **vol. 10**, no. 1, pp. 96, (2011).

[7]. J. Zhai and A. Barreto, *"Stress Detection in Computer Users Based on Digital Signal Processing of Noninvasive Physiological Variables,"* International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 1355–1358, (2006).

[8]. F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, and M. Griss, *"Activity-Aware Mental Stress Detection Using Physiological Sensors,"* in Mobile Computing, Applications, and Services, Berlin, Heidelberg, pp. 282–301, (2012).

[9]. A. de Santos Sierra, C. Sanchez Avila, J. Guerra Casanova, and G. Bailador del Pozo, *"A Stress-Detection System Based on Physiological Signals and Fuzzy Logic,"* IEEE Trans. Ind. Electron., **vol. 58**, no. 10, pp. 4857–4865, (2011).

[10]. P. Siirtola and J. Röning, *"Comparison of Regression and Classification Models for User-Independent and Personal Stress Detection,"* Sensors, **vol. 20**, no. 16, Art. no. 16, (2020).

[11]. 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, New York, NY, USA, pp. 400–408, (2018).

[12]. D. Makowski *et al.*, *"NeuroKit2: A Python toolbox for neurophysiological signal processing,"* Behav. Res. Methods, **vol. 53**, no. 4, pp. 1689–1696, (2021).

TÓM TẮT

Phát hiện căng thẳng trong quá trình lái xe sử dụng các tín hiệu sinh học và học máy

Căng thẳng (stress) là một vấn đề ảnh hưởng tới cả sức khỏe thể chất và tinh thần, gây ra các trạng thái cảm xúc tiêu cực. Căng thẳng có thể làm suy yếu khả năng nhận thức và xử lý tình huống của người lái trong việc lái xe an toàn. Do vậy, việc phát hiện và đánh giá mức độ căng thẳng đóng một vai trò quan trọng trong việc cải thiện sự dễ chịu và nâng cao trải nghiệm điều khiển phương tiện của người lái. Với bộ dữ liệu AffectiveROAD, bài báo này đề xuất một phương pháp phân loại mức độ căng thẳng dựa vào các tín hiệu sinh học. Những tín hiệu này được hiệu chuẩn thời gian và tiền xử lý để trích xuất các đặc trưng phù hợp trong chu kỳ 5 giây. Dựa vào các đặc trưng thu được, các mô hình học máy được huấn luyện để phân loại trạng thái stress thành 5 mức độ khác nhau. Kết quả kiểm nghiệm cho thấy độ chính xác đạt đến 94% với mô hình Random Forest (RF) khi sử dụng 7 đặc trưng quan trọng nhất từ các tín hiệu HR, EDA, TEMP, và tới 99% khi tích hợp thêm kỹ thuật chồng chập dữ liệu.

Từ khóa: Phát hiện căng thẳng; Wearable sensors; Trích xuất đặc trưng; Học máy; Random forests.