A Multi-Sensor Fusion Method For Stress Recognition

Leonardo Alchieri[1] [2], Nouran Abdalazim[1], Lidia Alecci[1], Silvia Santini[1], Shkurta Gashi[1]

***Abstract*— Modern wearable devices have enabled continuous and unobtrusive monitoring of human's physiological data such as heart rate. Adequate preprocessing of such data and application of machine learning enables automatic recognition of human behavior. In this paper we propose a multi-sensor approach that leverages physiological data collected using wearable sensors to detect stress. Our approach uses a temporal and sensory fusion methodology to leverage capabilities of multiple single-sensor models. To evaluate our approach, we use the SMILE dataset, which has been collected from 45 participants over 8 days. Our results show that the electrocardiogram and skin temperature's model fusion achieves an F1-score of 61.84% and an accuracy of 56.19% on the test set, which are higher than baseline classifiers. Our findings show the challenging nature of stress detection using physiological data and open up novel opportunities for further research. This paper is part of the stress detection challenge organized at the EMBC 2022 conference and presents the results of *MUSE-USI* team.**

I. INTRODUCTION

Stress is one of the critical daily defiances that can hamper the quality of life. People experience stress on a daily basis, e.g., due to their routine and/or work environment [1]. Stress can be classified into: *positive stress*, that can help individual be geared up and lead to increases in daily performance, and *negative stress*, often resulting from long periods of stressful events, and can lead to a chronic state, that impacts mental and physical health [2]. Automatic stress detection techniques have been proposed to help people enhance their quality of life. Nowadays, such techniques play a crucial role in managing one's stress level and reducing health risks [3].

Physiological signals are considered of uttermost importance for stress state detection, due to the strong correlation between stress and the autonomic nervous system [2], [4]. Some of the widely used physiological parameters are: blood volume pulse (BVP), photoplethysmography (PPG), electrocardioGram (ECG), galvanic skin response (GSR), also known as electrodermal activity (EDA), and skin temperature (ST), that can be captured in a continuously and unobtrusively using wearable devices [1].

The objective of this paper is to propose a multi-sensor fusion method to detect stress, with a focus on the simplicity and explainability of results. The implementation of our approach is publicly available on GitHub (`https://github.com/LeonardoAlchieri/MUSE`).

II. RELATED WORK

S25tress detection techniques can be grouped in *unimodal* and *multimodal*, as discussed in [5]. Unimodal approaches, such as [6], [5], [7], [8], use one sensor modality to estimate stress. Rashid et al. [6], for instance, propose a hybrid approach using a convolutional neural network (CNN) to distinguish between stress and non-stress states using BVP data collected with wristbands. Greco et al. [7] present a stress detection approach based on extensive preprocessing of EDA traces and support vector machine classifier. Authors in [5] investigate the stress detection performance, using single modality sensor in comparison with combined modalities, to economise computational requirements. In [8], an Artificial Intelligence-based fuzzy assisted Petri net method is proposed for stress detection, based on Heart Rate. These approaches show the feasibility of using physiological signals for stress detection, which is the goal of our paper.

Multimodal methods, on the other hand, leverage the information from two or multiple data sources for stress detection [9], [10], [11], [12]. Gil et al. [9], for instance, propose a binary stress detection system that relies on CNNs, along with inertial signals such as accelerometer (ACC) and physiological signals such as BVP, EDA, ECG, electromyogram (EMG) and respiration from two wearable devices. Wu et al. [11] leverage transfer learning along with handcrafted and deep features extracted from EDA, PPG and ST to detect stress. In [12], a real time binary stress detection system is proposed, where heart rate variability (HRV) and GSR features, collected using the wrist-worn devices, were employed, along with a voting and similarity-based fusion (VSBF) method. In [10], authors utilize unlabelled physiological and behavioral data to support the robustness of the stress classification problem, in a semi-supervised framework for stress detection, consisting of data augmentation, autoencoders and consistency regularization.

Building upon the work mentioned above, we also leverage multiple sensors to recognize stress. In contrast, we focus on understanding the most effective strategy to fusion sensor data and the overall impact of a sensor on stress detection.

III. DATA ANALYSIS

In this section, we describe the dataset used and our data analysis approach for stress detection.

A. Dataset

To evaluate our approach, we employ the **SMILE** (momentary stress labels with ECG, GSR, and ST data) dataset presented in [10]. It includes physiological data collected from 45 healthy participants (39 female, 6 male). The SMILE dataset was provided as part of the *EMBC 2022 Workshop and Challenge on Detection of Stress and Mental Health Using Wearable Sensors*. Two devices were used to collect

[1]Università della Svizzera Italiana (USI), Lugano, Switzerland
[2]leonardo.alchieri@usi.ch

the physiological data. The first device is *Chillband*[1] (IMEC), which collects GSR, ST and ACC; and the second device is *Health Patch* (IMEC)[2], with monitors for ECG and ACC. Participants reported their stress levels, on a scale from 1 (not stressed) to 7 (extremely stressed), several times during the day. The self-reported stress score was then assigned to 60 minutes of physiological data before the reported label.

1) Extracted Features: The challenge organizers provided features, described as follows. The features were extracted on a window of 5 minutes, with 4 minutes of overlapping, over the 60 minute interval. This results in a time series of 60 values. The extracted features of the ECG signal are: mean heart rate, heart rate cycle, low and high frequency signal, and their ratios, the ratio between low and very low frequency, the root mean square error of R-R differences and the standard deviation of the R-R intervals; the GSR are: the mean, the signal power of the phasic component, the response rate, the second difference, the response, the magnitude, the duration and the area and for ST: the mean, the standard deviation, the median and the slope of the fitted linear regression. Thus, in total, the dataset is composed of 2070 labels, for which there are 20 timeseries of length 60, each corresponding to a different feature. Another test set of 986 labels was used to benchmark the best method for the challenge, even though no ground truth was provided directly.

2) Stress Labels: We binarized the labels by considering the 1 to 7 self-reported stress score as the *stress* class and 0 as the *no stress* class, as suggested in [10]. With this paradigm, the dataset was balanced, with about 52% of the data samples in the positive class (stress) and 48% in the negative class (not stressed). The dataset authors also provided a set of features extracted using autoencoders, called *deep features*. Since in this work we focused on the explainability of the approach, we did not use such features.

B. Data Exploration

We explore the presence of missing values in the dataset and the relationship between stress and extracted features.

1) Missing Values: The dataset contains two types of missing data. The first group refers to the missing values of some of the features (27.25% of the cases), which were imputed by dataset authors with 0s. A smaller case, i.e., around 0.05% of the whole dataset, also presented some instances where, in a whole feature's sequence, one or more values were empty. These were solved with a simple average imputation, over the other points in the affected timeseries.

2) Correlation: We then performed correlation analysis to investigate the relationship of features to stress labels and to understand which timestamp has higher association with the label. To this goal, we explored the correlation using Pearson's ρ, Kendall's τ and Spearman's ρ [13]. We found no significant difference between the correlation measures, for this reason we report only the first one. Figure 1 shows the correlation between the features and stress labels. We observe that the ECG features have the highest correlation with respect to the stress label, which is in line with results reported in [14], with some ST features as well. The GSR features do not show any correlation with the stress label.

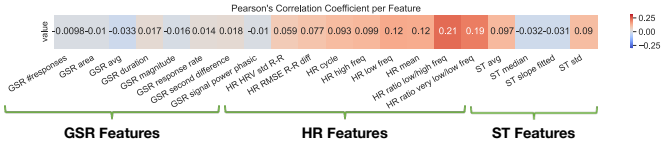

Fig. 1: Pearson's Correlation coefficient for all of the features. The p-value threshold was set to 0.05.

C. Classification Procedure

To recognize stress from the physiological data, we tested different machine learning algorithms. We implemented the approach using `Python` and the `Scikit-learn` library [15].

1) Timestep Selection: As mentioned in Section III-A, for each label, and each feature in it, a 60-value timeseries is given. Most work in the literature, e.g., [16], however, uses shorter windows; and, as described through saliency maps in [10], in the given dataset times closer to the labels should be more important. A decision to **reduce** the number of time values was thus taken. As such, instead of using all of the 60 window-averaged values, a shorter window of **10 minutes**, before the label taken, shall be used throughout the current analysis. We also experimented with other window sizes, such as, e.g., 30 minutes, and noticed that the 10 minute window showed the highest results.

2) Classifiers: For all sensors and modalities, i.e., multimodal or unimodal, we implemented a series of "classical" Machine Learning models: Gaussian Process (GP), Quadratic Discriminant Analysis (QDA), Support Vector Machine (SVM), Gaussian Naïve Bayes (NB), K-Nearest Neighbours (KNN), Decision Tree (DT) and some ensemble variations, like XGBoost (DT-XG), AdaBoost (DT-Ada) and Random Forest (RF).

3) Single-sensor models: We first investigated the use of only one sensor modality to recognize stress, to which we refer to as *unimodal* approach. However, since the data provided consists of 2 dimensions for each label, i.e., each feature is assigned to a timeseries, a few important considerations had to be made: most traditional machine learning algorithms cannot deal with multi-dimensional inputs. For simplicity, single label's data can be defined as $x \in \mathbb{R}^{T \times F}$, where T is the length of the timeseries and F is the number of features, e.g., 8 for the ECG data. The approach used to solve this, which we called *feature unravelling*, assigned the ground truth to each of the 10 values in the timeseries; in this case the number of total labels over which a classifier is trained is going to be 10 times more; $x \mapsto \left\{x'_j \in \mathbb{R}^F\right\}_{j=1}^{T}$. It is assumed fair given the shorter timestep selected: longer choices, e.g., 60 minutes, might hinder this approach.

[1] `https://drupal.imec-int.com/sites/default/files/2017-02/CHILL%20BAND_0.pdf`

[2] `https://www.imec-int.com/sites/default/files/imported/HEALTH%2520PATCH.pdf`

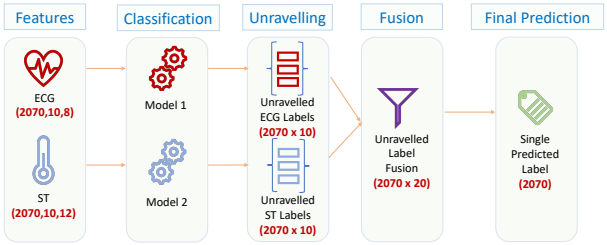

Fig. 2: Overview of our multi-sensor approach.

We then performed feature selection using the mutual information [17], which measures the mutual dependence of two variables. We did not observe any performance improvement when adopting it. Accordingly, we used all the features in the classification pipeline.

4) Multi-sensor models: We then investigated the use of multiple sensors to detect stress. The other objective laid out, in this setting, was to aid the explainability of the detection model. Based on this, we trained N independent models, where N is either 2 or 3 sensors, over the "unravelled" labels; and then used a *fusion method* to join the $N \cdot T$ predictions, i.e., the 10-values in the timeseries per sensor. Figure 2 shows the mutlimodal approach applied in more detail. Different combinations of ML sensor models, as well as fusion methods, both ML-based or not, were tested. As for the fusion modality, it was also tested whether using *probabilities*, i.e., instead of a label, the confidence that the classifier gives for the prediction, or directly label predictions could give different results. Feature selection was tested here was well, but without interesting results.

5) Evaluation Procedure: All models were trained and tested on the "train set", as provided for the competition, using a **10-fold cross validation** procedure, with **accuracy** as performance metric, and standard errors with 68% confidence. We further evaluated the performance of our approach on the test set through the automatic platform `CodaLab` (`https://codalab.org/`). This was used only to confront the best models, identified through the cross validation procedure, as to avoid introduction of bias, e.g., overfitting on the test set [18].

For the model identified as best, through the test set accuracy, as obtained in Section IV, a description of its results shall be provided, based on simple explainable AI practices [19]. Namely, which features, for the two signals, bear the most importance; and which timestep and sensor, during the model fusion, is more important. Both were obtained using an accuracy-based *feature permutation* [20] metric. We also calculated a confusion matrix, averaged over the cross validation folds, to identify how the model performs.

IV. RESULTS & DISCUSSION

1) Single-sensor models: As mentioned before, different techniques to deal with the dimensionality of the input data were tested. Table I shows the accuracy of the classifiers in comparison to the baseline, using the unravelling technique.

ML Model \ Sensor	ECG	GSR	ST
Gaussian Process	$\mathbf{58 \pm 3}$	40 ± 3	$\mathbf{60 \pm 3}$
SVM	58 ± 3	44 ± 4	59 ± 4
Naïve Bayes	54 ± 3	49 ± 6	57 ± 5
AdaBoost	54 ± 3	$\mathbf{50 \pm 3}$	56 ± 3
KNN	50 ± 2	49 ± 1	51 ± 2
QDA	51 ± 2	48 ± 6	59 ± 4
Uniform Random Baseline	*50 ± 2*		
Biased Random Baseline	*52.75 ± 0.06*		

TABLE I: Accuracy (%) for some classifiers and sensors (single-sensor)

Single-Sensor Model	SVM		GP	
Fusion Technique	CV	Test	CV	Test
Average	59 ± 3	51.52	60 ± 3	54.67
Gaussian Process	$\mathbf{60 \pm 3}$	52.74	$\mathbf{61 \pm 3}$	52.33
SVM	$\mathbf{60 \pm 3}$	53.14	59 ± 3	51.42
AdaBoost	59 ± 3	**54.56**	60 ± 3	51.72
QDA	55 ± 3	54.16	57 ± 5	**56.19**
Uniform Random Baseline	*50 ± 2*			
Biased Random Baseline	*52.75 ± 0.06*			

TABLE II: Accuracy (%) for combination of ML models and Fusion Techniques for the ECG+ST multi-sensory approach. Test accuracy is with two decimals, as provided by CodaLab. All others are rounded according to their standard error.

The accuracy level is modest for all models, even though above a random baseline, but nonetheless in line with the results obtained by [10]. Some classifiers performed better than others, e.g., SVM and GP, with somewhat consistency among sensors. As for sensors, ECG and ST have similar results, while GSR models are not statistically different than the random baseline, as might have been expected from the correlation analysis.

2) Multi-sensor models: In the multimodal apprpach, we used only the GP and SVM models, as they have shown to perform best for the single sensor experiments. As for the combination of sensors, while all were tested, meaningful results were possible only when using ECG and ST together, aligned with what obtained for the single-sensor approaches.

A decision to use the probability predictions, for the fusion phase, was taken, as opposed to the label predictions. During our tests, no discernible difference could be found between the two, when using cross-validation over the train set. However, from some performance benchmarks over the test set, was found that the probability predictions did indeed perform slightly better, given the same initial conditions.

Table II reports the classification results using the multi-sensor fusion technique applied on the prediction probabilities. In this case, since the results over the cross-validation technique were, most of the time, not statistically significant with one another, the accuracy over the test set, is also reported. The models used for the ECG and ST features were always the same together; some tests with different combinations were constructed, but at best the same results were obtained, and at worst a decrease in performance. From this analysis, the best model can be identified as the one which uses a Gaussian Process for the first phase, over both ECG and ST features, and a Quadratic Discriminant Analysis

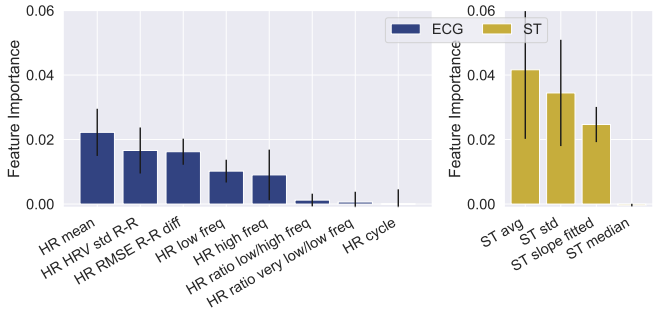

Fig. 3: Feature importance (over 10-fold cross validation), for the two models that make up the multi-modal approach, i.e. SVM for ECG and SVM for ST. Confidence intervals at 68% confidence are also shown.

(a) Heatmap with the feature importance, calculated with feature permutation.

(b) Confusion matrix (10-fold cross validation).

Fig. 4

for the fusion modality.

3) Best model analysis: Figure 3 presents the feature importance for the ECG and ST sensor using the Gaussian Process model. The most important features for the ST are slope fitted, the average of the signal and the standard deviation. As for the ECG data, all but 3 have high importance, i.e., the two ratios, between low and high and low and very low frequencies in HR signal, the high frequency values, the HR cycle and the standar deviation of the R-R peaks. The others were not statistically different than 0.

Figure 4a shows the importance of each timestamp for ECG and ST, from the fusion model. The fusion method, over the train set, considers the ST data as more important, for almost all timestep. This means that, according to the QDA fusion method, the ST-model predictions are more discriminative. On the other hand, there is no pattern for what concerns time: this could be either due to the decision of using a shorter window, and as such all values are already somewhat important, or the incapacity, on the fusion model side, in discriminating this. Future works could explore more in details this factor.

Figure 4b shows the averaged confusion matrix: the model can compute more accurately true labels (*stressed*) than false ones. The approach is capable of detecting one a person is stressed more often than when it is not: in a real world application, this behaviour could be desired for such a system.

V. CONCLUSIONS

In this paper, we proposed an explainable multi-sensor approach for binary stress detection. The proposed method relies on handcrafted features from the ECG and ST as well as classical machine learning algorithms. It achieves an F1-score of 61.84% and accuracy of 56.19% on the test set. In this work, we did not utilize the deep features provided by challenge organizers: investigating the impact of such features along with deep neural networks on the stress detection performance is an interesting future direction.

ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation (SNSF) project *PROSELF: Semi-automated Self-Tracking Systems to Improve Personal Productivity*.

REFERENCES

[1] A. Sano and R. W. Picard, "Stress recognition using wearable sensors and mobile phones," in *ACII*. IEEE, 2013.

[2] L. D. Sharma, V. K. Bohat, M. Habib, A.-Z. Ala'M, H. Faris, and I. Aljarah, "Evolutionary inspired approach for mental stress detection using eeg signal," *Expert Systems with Applications*, vol. 197, 2022.

[3] S. Gedam and S. Paul, "Automatic stress detection using wearable sensors and machine learning: A review," in *ICCCNT*. IEEE, 2020.

[4] P. Schmidt, A. Reiss, R. Duerichen, and K. Van Laerhoven, "Wearable affect and stress recognition: A review," *arXiv preprint arXiv:1811.08854*, 2018.

[5] R. Holder, R. K. Sah, M. Cleveland, and H. Ghasemzadeh, "Comparing the predictability of sensor modalities to detect stress from wearable sensor data," in *CCNC*. IEEE, 2022.

[6] N. Rashid, L. Chen, M. Dautta, A. Jimenez, P. Tseng, and M. A. Al Faruque, "Feature augmented hybrid cnn for stress recognition using wrist-based photoplethysmography sensor," in *EMBC*. IEEE, 2021.

[7] A. Greco, G. Valenza, J. Lázaro, J. M. Garzón-Rey, J. Aguiló, C. De-la Camara, R. Bailón, and E. P. Scilingo, "Acute stress state classification based on electrodermal activity modeling," *IEEE TAC*, 2021.

[8] Q. Lin, T. Li, P. M. Shakeel, and R. Samuel, "Advanced artificial intelligence in heart rate and blood pressure monitoring for stress management," *JAIHC*, vol. 12, no. 3, 2021.

[9] M. Gil-Martin, R. San-Segundo, A. Mateos, and J. Ferreiros-Lopez, "Human stress detection with wearable sensors using convolutional neural networks," *IEEE Aerospace and Electronic Systems Magazine*, vol. 37, no. 1, 2022.

[10] H. Yu and A. Sano, "Semi-supervised learning and data augmentation in wearable-based momentary stress detection in the wild," *arXiv preprint arXiv:2202.12935*, 2022.

[11] J. Wu, Y. Zhang, and X. Zhao, "Stress detection using wearable devices based on transfer learning," in *BIBM*. IEEE, 2021.

[12] S. A. Khowaja, A. G. Prabono, F. Setiawan, B. N. Yahya, and S.-L. Lee, "Toward soft real-time stress detection using wrist-worn devices for human workspaces," *Soft Computing*, vol. 25, no. 4, 2021.

[13] M. G. Kendall *et al.*, "The advanced theory of statistics." *The advanced theory of statistics.*, no. 2nd Ed, 1946.

[14] M. Elgendi and C. Menon, "Machine learning ranks ecg as an optimal wearable biosignal for assessing driving stress," *IEEE Access*, vol. 8, 2020.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *JMLR*, vol. 12, 2011.

[16] Y. S. Can, B. Arnrich, and C. Ersoy, "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey," *JBI*, vol. 92, 2019.

[17] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE mobile computing and communications review*, vol. 5, no. 1, 2001.

[18] V. Feldman, R. Frostig, and M. Hardt, "The advantages of multiple classes for reducing overfitting from test set reuse," in *ICML*. PMLR, 2019.

[19] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[20] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, 2001.