Stressalyzer: Convolutional Neural Network Framework for Personalized Stress Classification

Ramesh Kumar Sah[1], Michael John Cleveland[1], Assal Habibi[2] and Hassan Ghasemzadeh[3]

***Abstract*— Stress detection and monitoring is an active area of research with important implications for an individual's personal, professional, and social health. Current approaches for stress classification use traditional machine learning algorithms trained on features computed from multiple sensor modalities. These methods are data and computation-intensive, rely on hand-crafted features, and lack reproducibility. These limitations impede the practical use of stress detection and classification systems in the real world. To overcome these shortcomings, we propose *Stressalyzer*, a novel stress classification and personalization framework from single-modality sensor data without feature computation and selection. Stressalyzer uses only Electrodermal activity (EDA) sensor data while providing competitive results compared to the state-of-the-art techniques that use traditional machine learning models. Our single-channel neural network-based model achieves a classification accuracy of 92.9% and an $f1$ score of 0.89 for binary stress classification. Our leave-one-subject-out analysis establishes the subjective nature of stress and shows that personalizing stress models using Stressalyzer significantly improves the model performance. Without model personalization, we found a performance decline in 40% of the subjects, suggesting the need for model personalization.**

I. Introduction

Stress describes bodily reactions to perceived physical or psychological threats and is defined as the transition from a calm state to an excited state triggering a cascade of physiological responses [1], [2]. In the United States of America, around 77% of people suffer from headaches and insomnia related to stress. There has been a steady increase in the number of people suffering from stress-related issues each year [3]. Stress plays a critical role in many health problems, such as depression, anxiety, high blood pressure, heart attacks, and stroke [4]. Furthermore, stress influences a person's decision-making capability, attention span, learning, and problem-solving capacity [5]. Therefore, designing technologies that automatically infer moments of stress from sensor data is vital in providing appropriate interventions.

In this paper, we propose *Stressalyzer* - a novel stress classification and personalization framework from single-modality sensor data without feature computation and selection. We use a Convolutional Neural Network (CNN) with EDA data gathered using wearable devices for learning. Our trained model is competitive with other state-of-the-art methods in terms of performance. At the same time, our proposed approach does not suffer from many limitations inherent in earlier studies, such as high computing complexity, complex system design, and the burden of feature engineering and selection. All the experiments and results presented in this paper are fully reproducible with the code and data made publicly available[1].

A. Related Work

Traditionally, multiple sensor modalities or data streams such as heart rate variability (HRV), body acceleration (ACC), skin temperature, electrodermal activity (EDA), blood volume pulse (BVP), respiration rate, and electrocardiogram (ECG) are used to compute a large number of statistical and structural features to train stress classification models. In [7], authors computed 67 features from 7 sensor modalities to train a stress classification model with the best accuracy of 92.28%. Authors in [4] used deep neural networks (DNN) and 40 engineered features to achieve an accuracy of 95.21%. In [5], authors used statistical features and representations learned by a deep learning model trained on EDA data as features to train Bayesian and Tree-based stress classification models with an accuracy of up to 92%. Motivated by the results from [7], authors in [8] computed 195 features in time, frequency, entropy, and wavelet domain from chest and wrist EDA data to train the XGBoost algorithm with the highest f1-score of 0.89. Furthermore, several studies explored using sensors other than electrodermal activity for stress classification. In [9], authors used data from the built-in smartphone accelerometer sensor to identify activities that corresponded with stress levels and achieved an accuracy of 71%. Additionally, in [10], data from a commercial smartwatch were used for binary stress classification with an accuracy of up to 83%.

B. Contributions

Using data from multiple sensors or channels and computing a large number of features to train machine learning algorithms for stress classification has several disadvantages: (i) using many different sensors makes the system design complicated, expensive, and difficult-to-deploy in everyday living situations; (ii) a larger number of sensors translates

[1]Ramesh Sah is a Computer Science graduate student at Washington State University, Pullman, USA. ramesh.sah@wsu.edu. Dr. Michael Cleveland is an associate professor in the Department of Human Development at Washington State University, Pullman, USA. michael.cleveland@wsu.edu

[2]Dr. Assal Habibi is an Assistant Research Professor of Psychology at the Brain and Creativity Institute at University of Southern California, USA. ahabibi@usc.edu

[3]Dr. Hassan Ghasemzadeh is an associate professor in the College of Health Solutions at Arizona State University, Tempe, USA. hassan.ghasemzadeh@asu.edu

[1]*https://github.com/rameshKrSah/WESAD-stress-classification-personalization*

into a larger amount of energy to operate the sensors. Therefore, reducing the amounts of sensors improves the battery lifetime of the wearable system; and (iii) the model performance relies on the computed features, and choosing the right features requires domain knowledge. Besides, computing many complex features can make the classification algorithm less efficient in terms of run-time, energy, and memory. Furthermore, feature selection becomes an important step to select the most meaningful features and adds an extra processing step to an already complex machine learning pipeline.

Motivated by these drawbacks of multi-modal feature-based stress classification algorithms, our stress classification and personalization framework *Stressalyzer* uses $1D$ CNN with single-channel raw EDA sensor segments as inputs and learns the representations needed for classification without feature computation and selection at training time. Our primary objective is to implement a stress detection and classification system using only the EDA data. The secondary goal is to explore the personalization of stress models. Perception and effects of stress are subjective. The same external stimuli can have a varying degree of effect on different individuals regarding stress and emotional arousal. Hence, we also investigate whether stress detection algorithms require personalization or not.

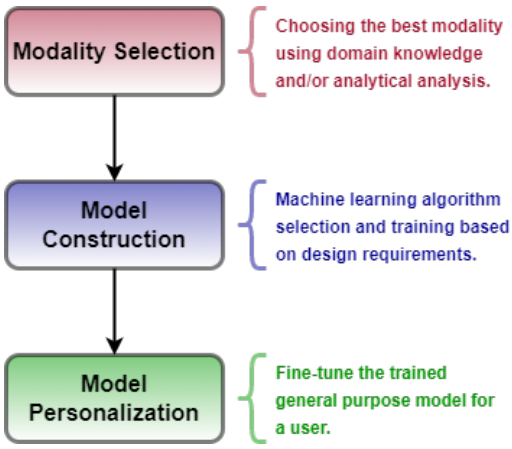

Fig. 1. Stressalyzer framework for stress classification and personalization.

II. Stressalyzer Design

We propose *Stressalyzer*, shown in Figure 1, for stress classification and personalization using time-series sensor data gathered with wearable sensors. Our approach has 3 main steps including *modality selection*, *model construction*, and *model personalization*. Modality selection tackles the problem of choosing the best modality for stress classification. Heuristics based on domain knowledge and/or analytical methods can guide modality selection. In model construction, we select and train an appropriate machine learning algorithm for stress classification using the data from the selected modality. The design requirements of the system guide model selection. The purpose of model personalization is to fine-tune a general stress classification model, obtained during model construction, for a particular user so that the model does not suffer from a drop in performance typically observed in machine learning applications [13], [14].

A. Modality Selection

The physiological changes associated with stress and emotional arousal are governed by the Autonomic Nervous System (ANS). The ANS has two main components: the Sympathetic Nervous System (SNS) and Parasympathetic Nervous System (PNS). Both SNS and PNS affect different bodily functions depending on a person's emotional state. During stressful episodes, a person undergoes many physiological changes such as increased heart rate, increased breathing rate, muscle tension, and sweating due to the changes induced by the SNS and PNS autonomic branches. Among all physiological signals, Electrodermal Activity (EDA), which is the measure of skin conductance, is the most sensitive to stress level due to the high correlation between EDA and SNS [12]. Consequently, EDA lends itself as a particular measure for stress because the SNS exclusively innervates skin sweating. Many earlier works have also used EDA, alone or together with other modalities, for stress classification [5], [7], [9]. Because our goal here is to design a one-channel stress monitoring approach, we select the EDA modality for training and personalizing machine learning model for stress classification.

B. Model Construction

Figure 2 shows a general architecture for a CNN model used in Stressalyzer. A convolutional neural network (CNN) is a representation learning algorithm capable of learning local dependency and scale invariance in the input. The associations between input and outputs are learned directly from raw data without feature computation.

In a convolutional layer, the convolution operation is used between the input and a weight matrix or kernel to assemble complex features by learning smaller and simpler features. The convolution operation slides the kernel over the input and computes the dot product of the kernel and the portion of the input segment to create a feature map. Having multiple kernels in a convolutional layer gives multiple feature maps, where each kernel looks for a specific type of concept in the input. Complex feature maps are extracted by stacking multiple convolutional layers on top of each other. The kernel at earlier layers represents low-level features, and the kernel in higher levels corresponds to more complex features. Therefore, the output of a convolutional layer is multiple feature maps corresponding to different kernels, and the kernel can be thought of as a feature extractor and the feature map as feature values. The feature maps are aggregated using some pooling operation to obtain a single feature vector. These feature vectors are the output of the convolutional stack and are used as input for the dense stack where the association between input sensor segments and output classes are learned. The convolutional stack enables the network to learn the hierarchy of features from the raw sensor data and

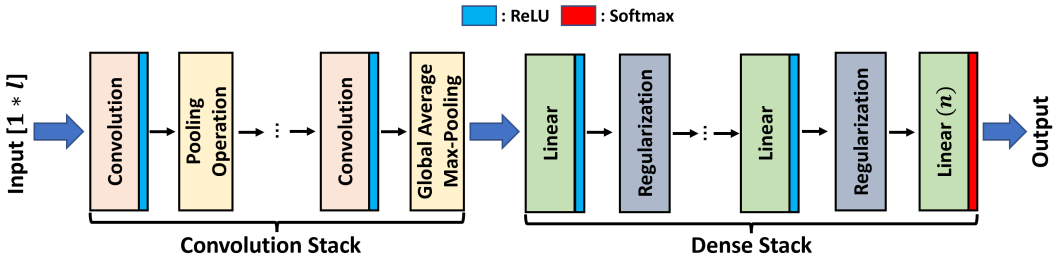

Fig. 2. General architecture of a $1D$ convolutional neural network used for time-series data. Input sensor segments of length l are passed through a convolutional stack that learns sensor data representations. These representations act like features for the dense stack and enable the dense layers to learn the associations between the input and output classes. The recognition layer handles the prediction of n output classes.

enable learning without traditional feature computation, and selection [15].

In our analysis, due to the time-series nature of the EDA data, we use $1-$dimensional CNN architecture composed of two convolutional layers with 100 filters and a kernel size of 5 and 10 respectively. Convolution layers are followed by a global max-pooling layer and two fully connected layers with 128 and 64 neurons. We also have drop-out layers after each fully connected layer with drop-out values 0.3 and 0.2. The output layer has Softmax activation, and all other layers have ReLU [11] activation.

C. Model Personalization

Perception and effects of stress are subjective. The same external stimuli can have varying degrees of effect on different individuals in terms of stress, and emotional arousal [1]. Furthermore, machine learning models trained on a general dataset often suffer from performance decline when used in personal settings [13], [14]. To deal with this challenge, we use an online learning method to personalize the stress model to a specific user. In the online learning scenario, a general machine model M_1 is retrained on data obtained from the user while the model is in use. The model is retrained until the performance of the personalized model, (M_2), on the user data is at an acceptable level. The model personalization approach used in *Stressalyzer* is shown in Figure 3. The goal of model personalization is to fine-tune a general model to the characteristics of a user. Personalization is a repeating process that continues until a given performance criterion is met.

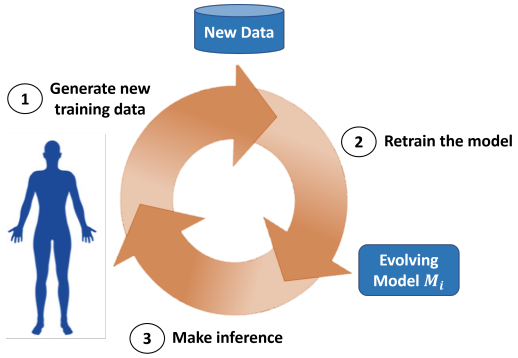

Fig. 3. Online learning scenario for personalization of machine learning models.

III. Validation and Results

A. Dataset

The Wearable Stress and Affect Detection (WESAD) dataset [7] is a publicly available dataset with ECG, EDA, BVP, respiration (RESP), skin temperature (TEMP), and motion (Acceleration) (ACC) sensor data obtained from the RespiBan (chest-worn) and Empatica E4 (wrist-worn) devices. The dataset was collected from 15 subjects (3 females) in a laboratory setting, and each subject experienced three main affect conditions: baseline or normal (neutral reading), stress (exposed to Tier Social Stress Test (TSST)), and amusement (watching funny videos). In our analysis, we only use the EDA data from the Empatica E4 sampled at 4Hz. Approximately the length of the stressed condition was 10 minutes, amusement 6.5 minutes, and baseline situation was 20 minutes.

B. Segmentation and Normalization

We segment the EDA data for the three affective states into 60 seconds overlapping windows with 50% overlap between consecutive segments. We settled on the window size of 60 seconds because of available literature that has also used 60 seconds window size for the WESAD dataset [4]–[7]. Before segmentation, we normalize the data for each subject using the min-max normalization to spread the data in the range of $[0, 1]$. After segmentation, we obtain 564 samples for the not-stressed class, 311 samples for the stressed class, and 165 samples for the amusement class. Our analysis has not used any method to deal with class imbalance, and machine learning models are trained on the imbalance data for the worst-case analysis.

C. Hyperparameters Tuning and Performance Metrics

The hyperparameters used in the CNN were selected after a random search over a set of values. The CNN models were trained for 200 epochs with a batch size of 32 and a fixed learning rate of 0.001. Out of 876 samples in the dataset, 657 or 75% was included in the training set, and 219 or 25% belonged to the test set. For bi-affective state classification, data from the baseline (not-stress) and stressed classes were used to create the training and test sets. For tri-affective state

classification, data for all three classes: baseline, stressed, and amusement, was used to create the training and test sets. We use accuracy, precision, recall, and f1-score to measure the performance of the trained models on the training and test sets.

D. Stress Classification

First, we present the results for the bi-affective state classification, i.e., the binary case of stress Vs. not-stress classification. The trained CNN model achieved the best classification accuracy of 94.8% on the training set and 90.9% on the test set. Table I, shows the value of other performance metrics.

TABLE I

RESULTS FOR THE BINARY STRESS CLASSIFICATION CASE: STRESS VS. NOT-STRESS.

Dataset	Accuracy	Precision	Recall	f1-Score
Training Set	94.8%	0.96	0.88	0.92
Testing Set	90.9%	0.91	0.82	0.87

TABLE II

RESULTS FOR THE TERNARY STRESS CLASSIFICATION CASE: STRESS VS. NOT-STRESS VS. AMUSEMENT

Dataset	Accuracy	Precision	Recall	f1-Score
Training Set	85.1%	0.83	0.79	0.80
Testing Set	82%	0.82	0.72	0.76

In the second case, we consider the tri-affective state classification, a multi-class classification problem with 3 classes: stress, not-stress, and amusement. Table II shows the values of performance metrics for this case. Note that the performance of the CNN model has decreased in the tri-affective case compared to the bi-affective case. We suspect this is because the model does not have enough training samples to distinguish between the three classes.

TABLE III

AVERAGE RESULTS AFTER 10-FOLD CROSS-VALIDATION FOR BI-AFFECTIVE AND TRI-AFFECTIVE CASE.

	Dataset	Accuracy	f1-Score
Bi-affective	Training Set	93%	0.9
	Testing Set	90%	0.86
Tri-affective	Training Set	84%	0.79
	Testing Set	80%	0.75

TABLE IV

COMPARISONS OF OUR CNN MODEL WITH CURRENT STATE-OF-THE-ART METHODS.

Method	Model Type	Modalities	Channels	Accuracy (%)	f1-Score
[7]	Feature	All	10	93	0.9
[5]	Feature	Wrist EDA	1	91.6	-
[8]	Feature	Chest and Wrist EDA	2	–	0.89
[4]	Feature	All	10	95.21	0.94
Our's	**Data**	**Wrist EDA**	**1**	**92.85**	**0.89**

Furthermore, to account for the variance in performance, we conducted $10-$fold cross-validation for both cases of affective state classification. Table III shows the average classification accuracy and f1-score for bi-affective and tri-affective cases. Finally, we present comparisons of our results with other state-of-the-art works on stress classification with the WESAD dataset in table IV. WESAD dataset has the following modalities ACC, Wrist EDA, Chest EDA, TEMP, ECG, BVP, and RESP and all together 10 channels. All other compared approaches, details in I-A, computes statistical or representational features from sensor data to train machine learning models. Our method does not need to compute features, works with single channel raw wrist EDA sensor data, and automatically learns the mapping between inputs and outputs during train time. Using single channel EDA data will make our system simple and less resource-hungry in real-life applications. We found our proposed approach competitive with state-of-the-art methods with the added advantage of being data-driven without needing specialized domain knowledge and lower resource requirements.

E. Personalization Analysis

Stress is subjective, and the same external stimuli can have different effects on different individuals. To investigate the subjective nature of stress and answer whether we need personalized models for stress detection, we first present the results of our leave-one-subject-out (LOSO) analysis on the binary WESAD dataset. In LOSO analysis, data from one subject is removed from the training set and kept as the test set to evaluate the machine learning model trained on data from all other subjects. To quantify performance decline, we calculate the difference in the model's accuracy on the training and test set. If the difference is larger than $\delta = 5\%$, we conclude that the subject needs personalization. Figures 4 and 5 shows the classification accuracy and f1-score of the trained models on the test and training sets. The x-axis represents the subject whose data was not included in the training set and was used as the test set.

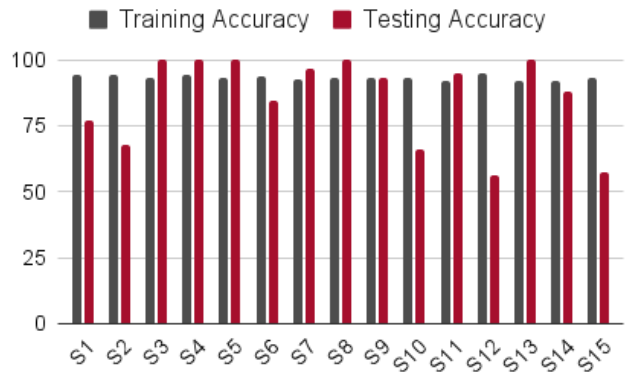

Fig. 4. Classification accuracy on the training and test sets for leave-one-subject-out analysis. The x-axis represents the subject whose data was not included in the training set and was used as the test set.

Out of 15 subjects, we observed a performance decline in 6 subjects, i.e., 40% of the total subjects. For the remaining subjects (60%), the trained model performed better or similar on the test set compared to the training set. In particular,

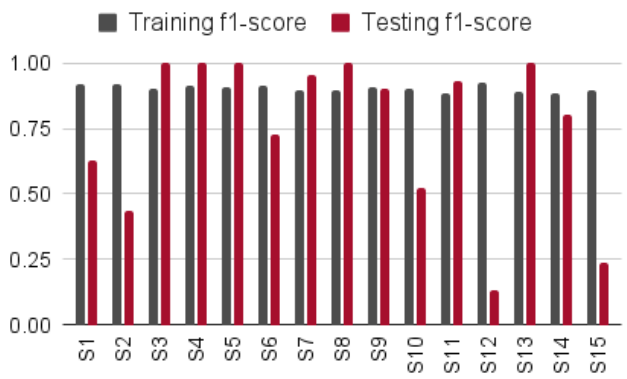

Fig. 5. f1-score on the training and test sets for the leave-one-subject-out analysis to investigate the subjective nature of stress. The x-axis represents the subject whose data was not included in the training set and was used as the test set.

we see performance decline for subjects *S1, S2, S6, S10, S12,* and *S15*. The decline in the model's performance on the test set is due to the inter-subject differences such as physical characteristics, emotional endurance, stress management skills, personality traits, and noise in the sensor data present in the dataset. For example, subject $S2$ was looking forward to stress conditions and was cheerful during data collection. Hence, for individuals with different responses to stress stimuli compared to the general outlook, the trained model failed to capture the personal traits of the individual. Next, to personalize stress models, we consider an online learning scenario. We ignore other details of the online learning paradigm, such as querying the user for labels and use the subject's data kept as the test set for retraining. Starting from 1 sample from the test set, we successively increase the number of samples used for retraining until the performance on the remaining subject data is greater or equal to that on the original training set. Table V shows the number of samples needed for retraining and the final test set accuracy. The personalized model performance on the test set increased significantly after retraining, establishing the subjective nature of stress and demonstrating the benefit of personalizing stress models.

TABLE V

TEST SET ACCURACY AND NUMBER OF SAMPLES NEEDED TO PERSONALIZE STRESS MODELS.

Subject	Before Personalization Accuracy (%)	Re-training Sample Size	After Personalization Accuracy (%)
S1	76.8	43	96.4
S2	67.9	56	83.9
S6	84.5	40	98.3
S10	66.1	52	98.3
S12	55.9	42	94.9
S15	57.4	43	93.4

IV. CONCLUSION

We presented the development and evlaution of *Stressalyzer*, a novel stress classification and personalization framework for single modality sensor data without feature computation and selection. We used the EDA modality and a CNN for stress classification. The trained model achieved a binary stress classification accuracy of 90%, and we also established the subjective nature of stress with our leave-one-subject-out personalization analysis. We found a performance decline in 6 (40%) of the subjects out of 15 total subjects when a general stress classification model was evaluated on the never-seen new subject data. We also showed that stress models could be personalized to achieve on-par or better performance for the new user using online learning. Our approach is competitive with the state-of-the-art methods while it does not suffer from disadvantages such as feature computation and selection, multi-modal input data, and complex system design, and is adaptable to include more sensor modalities for performance improvement.

REFERENCES

[1] Selye, Hans. The stress of life. 1956.
[2] Kasl, Stanislav V. Stress and health. Annual review of public health 5.1 (1984): 319-341.
[3] American Psychological Association. Stress in America: The state of our nation. Stress in AmericaTM Survey (2017).
[4] Bobade, Pramod, and M. Vani. Stress Detection with Machine Learning and Deep Learning using Multimodal Physiological Data. 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE, 2020.
[5] Aqajari, Seyed Amir Hossein, et al. GSR Analysis for Stress: Development and Validation of an Open Source Tool for Noisy Naturalistic GSR Data. arXiv preprint arXiv:2005.01834 (2020).
[6] Choi, Jongyoon, Beena Ahmed, and Ricardo Gutierrez-Osuna. Development and evaluation of an ambulatory stress monitor based on wearable sensors. IEEE transactions on information technology in biomedicine 16.2 (2011): 279-286.
[7] Schmidt, Philip, et al. Introducing wesad, a multimodal dataset for wearable stress and affect detection. Proceedings of the 20th ACM International Conference on Multimodal Interaction. 2018.
[8] Hsieh, Cheng-Ping, et al. Feature Selection Framework for XGBoost Based on Electrodermal Activity in Stress Detection. 2019 IEEE International Workshop on Signal Processing Systems (SiPS). IEEE, 2019.
[9] Garcia-Ceja, Enrique, Venet Osmani, and Oscar Mayora. Automatic stress detection in working environments from smartphones' accelerometer data: a first step. IEEE journal of biomedical and health informatics 20.4 (2015): 1053-1060.
[10] Gjoreski, Martin, et al. Continuous stress detection using a wrist device: in laboratory and real life. proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct. 2016.
[11] Jarrett, Kevin, et al. What is the best multi-stage architecture for object recognition?. 2009 IEEE 12th international conference on computer vision. IEEE, 2009.
[12] Ziegler, Michael G. "Psychological stress and the autonomic nervous system." Primer on the autonomic nervous system. Academic Press, 2012. 291-293.
[13] J. Xiao, L. Chen, H. Chen and X. Hong, "Baseline Model Training in Sensor-Based Human Activity Recognition: An Incremental Learning Approach," in IEEE Access, vol. 9, pp. 70261-70272, 2021, doi: 10.1109/ACCESS.2021.3077764.
[14] C. -Y. Lin and R. Marculescu, "Model Personalization for Human Activity Recognition," 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), 2020, pp. 1-7, doi: 10.1109/PerComWorkshops48775.2020.9156229.
[15] Rokni, Seyed Ali, et al. "TransNet: minimally supervised deep transfer learning for dynamic adaptation of wearable systems." ACM Transactions on Design Automation of Electronic Systems (TODAES) 26.1 (2020): 1-31.