Enhancing Stress Detection for Students: Exploring the Impact of Fine-Tuning and User-Specific Data Calibration in Deep Learning

Rashmi Alur Ramachandra* [1], Jayasankar Santhosh [2], Andreas Dengel [3], Shoya Ishimaru [4]
[123]Department of Computer Science, RPTU, Kaiserslautern, Germany,
[23]Smart Data and Knowledge Services, DFKI, Kaiserslautern, Germany,
[4]Department of Computer Science, Osaka Metropolitan University, Osaka, Japan

Abstract

This study presents a comprehensive investigation into stress detection among students, focusing on multiple levels of stress assessment. This research aims to shed light on the complexities of stress experienced in educational settings by utilizing a physiological sensing wristband to capture the multifaceted nature of stress responses. A user study was conducted to calculate the cognitive stress levels of a group of 25 participants by recording physiological signals on an Empatica E4 wristband. Along with the relaxed or non-stressed condition, the study employed a range of simple to complex arithmetic tasks designed to elicit three levels of response: 1) slightly stressed or easy level, 2) stressed or medium level, and 3) highly stressed or hard level. Upon the implementation of multiple deep learning models, FCN, ResNet, and LSTM models demonstrated promising outcomes in accurately categorizing the three different stress levels (easy, medium and hard). The models were trained using KFold and Leave-One-Participant-Out (LOPO) cross-validation techniques. To improve the prediction accuracy of LOPO, a fine-tuning or user-specific data calibration approach was utilized. This approach resulted in significant improvements in accuracy for LOPO, with the FCN model achieving a spike to 60% (F1=0.578), the ResNet model reaching 85% (F1=0.846), and the LSTM model achieving an impressive 91% (F1=0.911) accuracy

[1] rashmi.ar96@gmail.com
[2] jayasankar.santhosh@dfki.de
[3] andreas.dengel@dfki.de
[4] ishimaru@omu.ac.jp

for three-class classification. Leveraging the insights gained from the prediction outcomes, a prototype application was developed that effectively portrays the dynamic fluctuations in stress levels. This application incorporates a stress meter, allowing users to visually comprehend their stress levels, and it delivers customized alert messages to individuals based on their respective stress levels, ensuring timely support and intervention.

1 Introduction

Stress is commonly linked to a predominantly negative perception of individuals and is regarded as a subjective experience that can impact both emotional and physical well-being. Psychological stress has become a major and detrimental issue among young individuals, especially students, in the current society [2]. In the modern digital era, students are confronted with the necessity of attending multiple online classes and lectures, managing a considerable workload of assignments, and facing demanding exams. The cumulative effect of these responsibilities can lead to heightened levels of stress among students, significantly impacting their overall mental and physical well-being. Recognizing and addressing stress promptly is crucial for ensuring the student's welfare.

Historically, various physiological features such as electroencephalography (EEG), galvanic skin response (GSR), and electrocardiogram (ECG) have been extensively employed in the detection and assessment of emotions, mental workload, and stress [33][12][34][25][5][7]. These measures have proven to be valuable tools for evaluating the physiological responses associated with stress over the years [16]. By monitoring and analyzing these signals, researchers have been able to gain insights into the emotional and cognitive states of individuals, providing valuable information for stress detection. In the context of student's well-being, the utilization of these physiological features holds promise in identifying the presence and magnitude of stress experienced by students. By leveraging technologies that capture and interpret EEG, GSR, and ECG data, researchers can objectively assess the stress levels of students. This objective measurement can aid in early identification and intervention, enabling timely support systems to be put in place.

By incorporating these physiological measures into stress detection methodologies, educators and institutions can gain a more profound understanding of the stressors faced by students and tailor their educational environments accordingly. Additionally, this knowledge can inform the development of personalized interventions and coping strategies, promoting the overall mental health and academic success of students in the digital age. WESAD, a widely recognized dataset, played a significant role in exploring different affective states using two prominent sensing devices, namely Empatica E4 and RespiBAN. The dataset aimed to capture a comprehensive range of emotional experiences and physiological responses. Despite encompassing a broad spectrum of affective states, the evaluation process focused specifically on three distinct states: baseline, stress, and amusement [32].

In our study, we employed the Empatica E4 wristband as a data col-

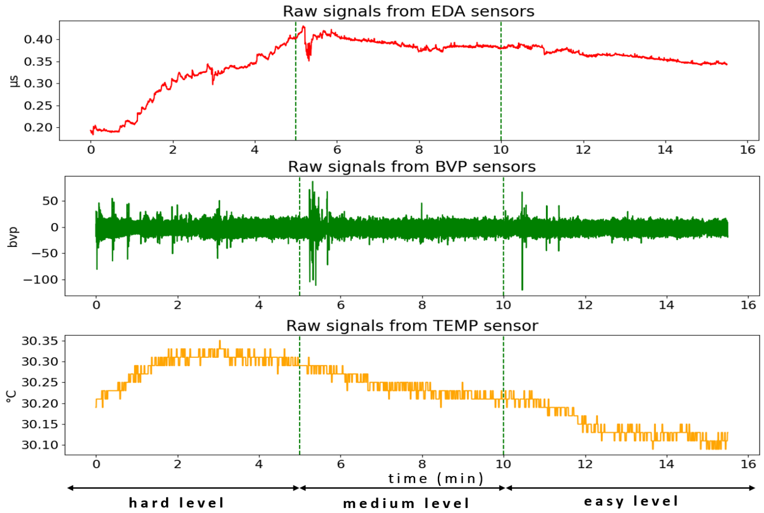


Figure 1: The raw physiological signals separated by different sessions of stress for a single participant.

lection tool to investigate and quantify the stress levels experienced by students. The primary objective of our research was to assess and analyze stress in students while they performed a series of tasks at different difficulty levels. Our focus was on understanding how students responded to varying levels of stress and whether there were discernible differences in their physiological reactions to different stress levels. To achieve this, we designed and conducted an experiment involving 25 participants and subjected them to three distinct stress conditions: 'easy', 'medium', and 'hard'. Figure 1 illustrates the variations in the raw sensor signals during different sessions in the study. To induce stress during the experiment, we utilized mental arithmetic tasks as stressors [22][37][26][27]. By asking the participants to solve arithmetic problems within specific time constraints, we aimed to create a stress-inducing scenario at various levels. Our research sought to contribute to the understanding of how students experience and respond to stress during task performance. By investigating their physiological responses under different levels of stress, we aimed to uncover patterns and variations that could shed light on the impact of stress on student's well-being and performance. The utilization of the E4 wristband and the collection of physiological data enabled us to capture objective measures of stress, complementing self-report measures and enriching our understanding of the participant's experiences. This approach allowed for a more comprehensive assessment of stress levels and provided valuable insights into the physiological manifestations of stress in students during task-based activities.

The key contributions of this study are:

- A comprehensive and well-designed experimental protocol with a well-established stressor, utilizing a range of tasks and stimuli that effectively induce varying degrees of stress in participants. This allows for controlled and consistent manipulation of stress levels during the experiment.

- An extensive analysis and comparison of multiple deep learning models for predicting stress levels and demonstrating that these models can be trained effectively to accurately predict stress levels based on physiological data.
- A comparative analysis of a general classification model to a user-adaptive model, implemented by fine-tuning or calibrating user-specific information.
- An application prototype interpreting the dynamic variations of stress levels in users, utilizing a stress meter, customized alert messages and serving as a tool for aiding in their understanding and monitoring of stress levels, ensuring timely support.

In a nutshell, our research emphasizes the significance of comprehending and monitoring stress levels in students. By employing a robust experimental protocol, analyzing physiological data with deep learning models, and developing an application for visualization and providing potential interventions, we contribute to the advancement of stress detection methods and their practical implementation.

2 Background and Related Work

Stress detection has a decade-long history in computer science. Detection and analysis of stress can range from uncomfortable sensors to comfortable wearable sensors, and from difficult situations to simple conditions and experiments. According to [3], human stress can be detected in two ways: 1) subjective questionnaires created by psychologists and 2) objective measurements that include physiological signals from wearable and non-wearable sensors. Some commonly used questionnaires are Perceived Stress Scale (PSS) [9], Daily stress inventory (DSI) [6] and Brief symptom inventory (BSI) [10]. Ideally, subjective measurements are less convincing without objective measurements, as objective measurements or physiological sensors help to detect human emotion using wearable devices placed on the participant's body without physical contact with them.

Cognitive stress has a strong connection to brain activity [11]. EEG is a non-invasive method of recording brain electrical activities. Medical advancements have resulted in the release of wearable EEG devices, which are easy to wear, head mounted, comfortable, and user-friendly. The dry, tiny and non-contact electrodes detect electrical charges due to brain activities. Similar to brain activities, the correlation between the human central system and the heart is used to measure human stress [1]. ECG is a non-invasive modality used to measure and monitor heart functions. RR intervals (rhythm to rhythm) and HR (heart rate) can also be extracted from ECG signals. ECG signals are observed as R-waves and ECG intervals are measured between them. Heart rate is its reciprocal. A study by Ahn et al. [1], demonstrates detection of stress based on time and frequency features of EEG and ECG. The study involved two stress conditions while performing two different tasks, Stroop color word and mental arithmetic tests, with 14 subjects. A traditional machine learning model, SVM, classified the two conditions with an accuracy rate of 87.50%.

Electrodermal activity (EDA), also known as the GSR, helps to read human skin variations through the sweat glands. There is a lot of information about the human state of mind that can be derived from these variations [40]. The spike in EDA can be due to physical activities like running, sleeping, standing or emotional activities like excitement, stress, fear, anger. Blood volume pulse (BVP) is another physiological feature that detects emotions [30]. In photoplethysmography (PPG), infrared light is transmitted through tissue and the absorption of this light by blood flowing through the vessels is measured [28]. It is also used to measure BVP, which is controlled by the heartbeat. In a stressful situation, the human body releases a lot of stress hormones that increase blood pressure. Gjoreski et al. [18] presented a multimodal stress classification framework incorporating EDA, BVP, HR, RR, and skin temperature (TEMP) collected from an Empatica E4 wristband from 26 participants. This continuous stress detection model used data from two scenarios, laboratory and real life. Stress detection in the laboratory was performed in two and three-level classifications, resulting in an accuracy rate of 83% and 72%. Real-life data achieved 76% and 92% accuracy respectively in no-context and context scenarios. Cho et al. [8] combined PPG and thermal imaging data to build a mobile stress monitoring system. This smartphone-based system recognizes stress instantly using neural networks.

Other physiological signals that contribute to detecting stress in humans are EMG (Electromyography) [23] and acceleration (ACC) data [14] [17]. EMG is a technique used to measure the health of muscles and nerve cells called motor nerves. EMG is controlled by the nervous system and depends on the physiological and anatomical characteristics of the human skeletal system, making it a complicated modality. A wide range of studies have proven the relationship between EMG and human stress. In a study presented by Ghaderi et al. [15], a multimodal stress detection method to understand driver's emotions is implemented using EMG, EDA and ECG. Using SVMs and kNNs, the three-level classification data has been generalized with 98% accuracy. An accelerometer measures the acceleration of the body using wearable devices or smartphones. It is used to detect stress in the course of daily activities by tracking body movements. Additionally, a gyroscope is a sensor that senses the angular velocity in x, y and z axis for detecting stress. In a study by Sysoev et al. [35], behavioral and contextual data collected in real-life scenarios was used to determine stress in a non-invasive way using smartphones. The data is collected from gyroscope, accelerometer, current stress level self-assessment and current activity type. Using only accelerometer data, an accuracy rate of 82.5% and 90.32% was achieved for daily activities and standing activity, respectively.

Tracking eye movement helps to measure where a person looks, which is called the point of gaze. These eye movements are converted to a stream of data that includes gaze point, gaze vector and pupil position. Human eye behavior is affected by different emotional situations [19]. A stress detection framework called StressClick, developed by Huang et al. [21], uses human gaze with mouse clicks of participants while performing mental arithmetic tasks. The system classified two stress scenarios using a Random Forest classifier with 60% accuracy.

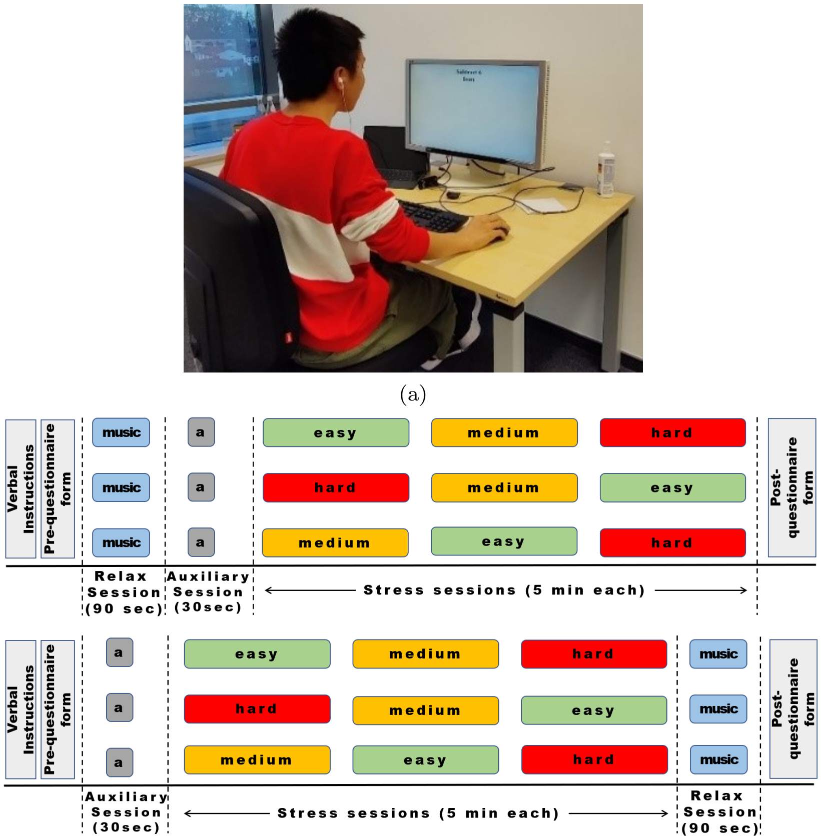
One of the major challenges while dealing with time series data like physiological signals is extracting meaningful information from them, also known as feature engineering. Most machine learning models fail when faced with large amounts of physiological data without feature extraction. Yan et al. [39], proposed baseline models for time series data using deep learning approaches. In this paper, the Multilayer Perceptron (MLP), Fully Connected Networks (FCN), and Residual Networks (ResNet) are implemented without any pre-processing or feature engineering involved. Similarly, in a paper by Dziezyc et al. [13], ten end-to-end multimodal deep learning architectures are presented that detect stress and other emotions without extracting features from raw physiological data. The study uses sensory information from four different datasets at a standard sampling frequency without pre-processing, to preserve all the information in the signals. In a recent study, Behinaein et al. [4] presented an end-to-end deep learning model for emotion recognition based on two publicly available datasets, WESAD [32] and SWELL-KW [24]. The research includes developing a neural network based on convolutional layers and multi-head transformers that are applied to ECG signals.

3 Methodology

We conducted an experiment involving 25 participants, 11 female and 14 male, all pursuing masters at the university and aged between 21 and 31. Our system recorded physiological signals of the participants while performing a series of tasks at different levels, as shown in Figures 2a and 2b.

3.1 Experimental Design

Before the start of the experiment, the experimental procedure and all the apparatus were described in detail to all participants. Students who agreed and signed the consent form could participate. We asked the participants to wear the Empatica E4 wristband on their non-dominant hand. Additionally, we requested all participants to keep their hand movements minimal, as we did not want hand gestures to add extra stress and increase variations. Subsequently, participants had to be engaged with the tasks coming one after the other for 20 to 25 min. The participants were mainly divided into 2 groups depending upon the order of relax and stress sessions; 1) relax followed by stress, 2) stress followed by relax. These two groups were further divided into 3 groups depending upon the stress levels; 1) easy-medium-hard, 2) medium-easy-hard, 3) hard-medium-easy as shown in Figure 2b. At the beginning and end of the relax and stress session, the participants were asked to tag/press the button on the E4 wristband. A red blink indicates each press. At the start of each session,



(a)

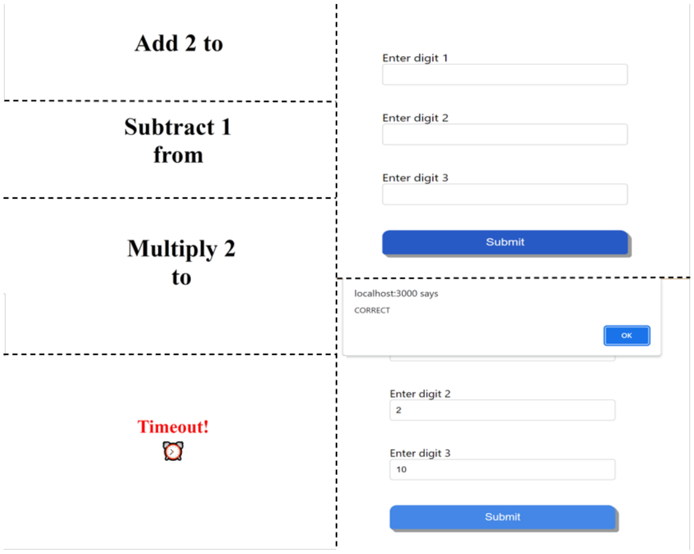
(b)

Figure 2: (a) An overview of the experimental setup. The participant is looking at numbers appearing on the screen in a series of mental arithmetic tasks. (b) The various sessions involved in the experimental setup

clear instructions about the upcoming session appeared on the screen. In this way, all the dos and don’ts of the experiment were explained to the participants.

Sessions:

1. Relax: In the relax session, the participants just sat in front of the monitor with a pleasant image on the screen and listened to soothing music for 90 sec.
2. Stress: The stress session included 3 levels of arithmetic tasks. All levels were 5 minutes each. A series of numbers were shown on the screen. Participants were asked to memorize the series and mentally “Add/Subtract/Multiply” a given number to each number in the series. For example, if the series shown is (5, 3, 1) and if asked to add 1, then the correct answer will be (6, 4, 2). In the following



Add 2 to

Subtract 1 from

Multiply 2 to

Timeout!

Enter digit 1

Enter digit 2

Enter digit 3

localhost:3000 says

CORRECT

Enter digit 2: 2

Enter digit 3: 10

Figure 3: Screenshots of experimental sessions. Easy level, Medium level, Hard level, 'timeout' message (left-top to bottom). Input field for participants to answer easy level, 'correct' message (right-top to bottom)

fields, they were required to input the revised series. If the input answer series is correct, an alert message "Correct" is popped on the screen and if not, an alert message "Incorrect" is popped. If they fail to input an answer, a "timeout" message appears on the screen as shown in Figure 3.

(a) In the easy level, participants solved addition problems. They had to memorize 3 numbers in series and input 3 answers in the same order.

(b) In the medium level, participants solved subtraction tasks. They had to memorize 4 numbers in series and input 4 revised answers in the same order. In addition to alert messages, participants heard a buzzer sound for correct, incorrect and timeout results.

(c) In the hard level, participants solved multiplication tasks. The total numbers in the series were increased to 5. In addition to alert messages and buzzer sounds for correct, incorrect and timeout results, participants heard a ticking clock sound when this session started.

3. Auxiliary session: Before the stress session began, participants had to pass through an additional task, in which a few circles appeared on the screen one after another. They were asked to catch these circles by clicking anywhere inside the circle. As we considered this session a complementary session which would allow the participants to concentrate on the upcoming arithmetic tasks, we did not use the wristband. This session lasted for 30 seconds.

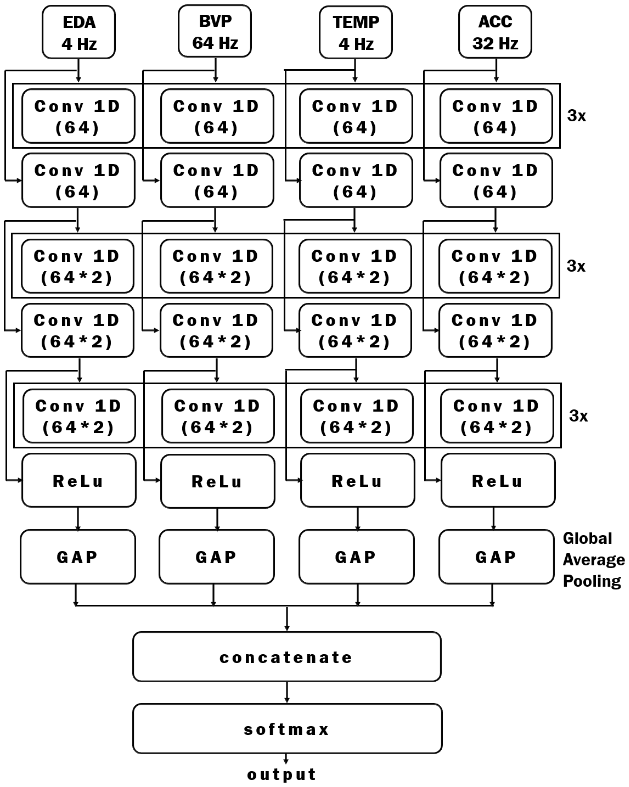


Figure 4: ResNet Architecture

In between each session, participants took a few seconds break. Besides the alert messages, the buzzer, and ticking sounds were used to induce stress or anxiety. Additionally, a questionnaire form was also administered at the start and end of the experiment, asking participants to answer a couple of questions about themselves and the experiment. The pre- and post-questionnaire forms had a Likert scale from 1 to 7 (where 1 being the least and 7 being the highest) and a drop-down menu to select their options.

3.2 Data Pre-processing

The physiological signals like EDA, BVP, TEMP, HR and 3-axis ACC data were collected from Empatica E4 with sampling frequencies of 4Hz, 64Hz, 4Hz, 1Hz and 32Hz respectively. The recorded EDA, BVP, TEMP and ACC signals were pre-processed and segmented using a sliding window of the length of 30 seconds without overlap. The input signals provided to Transformers and LSTM were resampled to a frequency of 4Hz, ensuring uniform frequency across all signals, while FCN and ResNet were

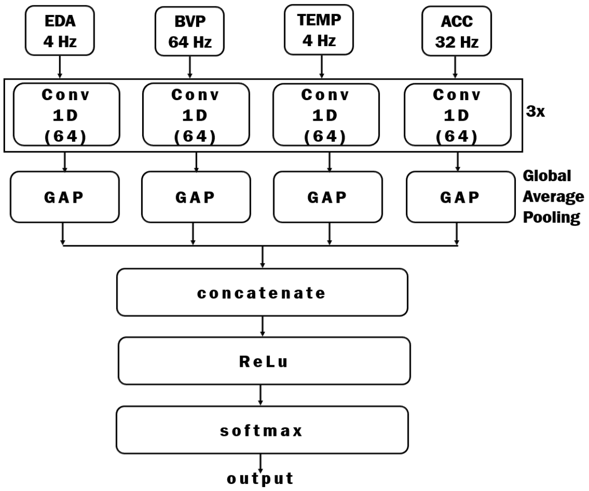


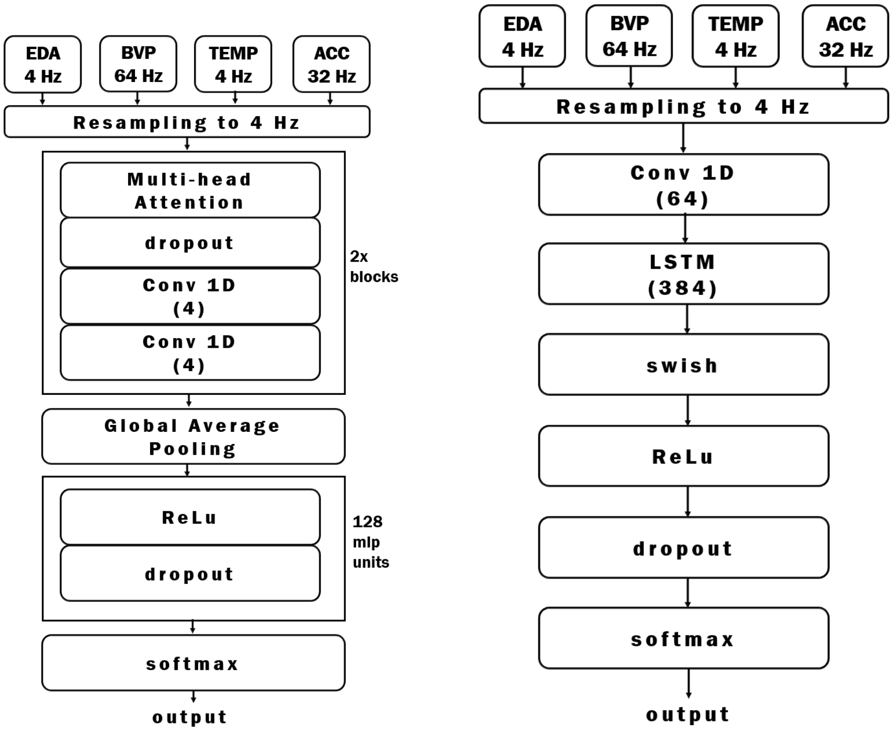
Figure 5: FCN Architecture

fed with signals at their original sampling frequencies. The signals were then standardized to have zero mean and unit variance before being fed into the networks. Our methodology employs an end-to-end approach, allowing us to extract valuable insights directly from the raw sensor data without the need for manual feature extraction. By bypassing the manual feature engineering step, our method enables a streamlined process that leverages the inherent information contained within the sensor data itself. This approach eliminates potential biases and limitations introduced by human-designed features, allowing for a more comprehensive and unbiased analysis of the data.

3.3 Model Architecture

Deep neural networks alleviate the necessity for feature engineering by acquiring high-level features within their hidden layers. This leads to a reduction in the complexity of the workflow and the amount of manual effort required, while simultaneously enhancing the likelihood of capturing relevant information.The collected data was analyzed using various models, including FCN, ResNet, Transformers and Long Short-Term Memory (LSTM). The diagram in Figure 4 shows the architecture of ResNet [39]. The network is highly dense, with multiple residual blocks and shortcut connections within each block. Our model contains three residual blocks where in each block, there are three successive convolutional layers with 64, 128 and 128 filters, respectively. The kernel size of convolutional layers also varies for each block. A global average pooling layer is incorporated to downsample the feature maps and capture the most relevant information. Towards the end of the architecture, fully connected layers are used to map the learned features to the desired output and the output layer provides the final prediction based on the processed physiological data. Figure 5 shows the architecture of the FCN model [39]. The FCN model consists of three convolutional blocks for each signal, followed by a global average pooling layer. The branches are concatenated and fed to one or more fully connected dense layers. Each convolutional layer applies a set of filters to the input time series, and the output of each layer is passed through a non-linear activation function such as the ReLU function. The convolutional layers output is processed by the fully connected layers to complete the final classification.

A Transformers model is applied to the preprocessed signals [36] [38]. These signals were first passed through the transformer block twice. The transformer block consists of a multi-head attention layer, a dropout layer, and two 1D convolutional layers. The transformer block is coupled with a pooling layer and dense layers with ReLu and Softmax as shown in the Figure 6a. Figure 6b illustrates the architecture of an LSTM model. Similar to Transformers, the inputs for LSTM [20] are sampled at a consistent frequency of 4Hz. The LSTM model comprises a 1D convolutional layer with a filter size of 64 and a kernel size of 7. This is followed by an LSTM layer and two dense layers, swish [31] and ReLu. Subsequently, the output is passed through a dropout layer and dense layer for further processing.



(a) (b)

Figure 6: (a) Transformers Architecture (b) LSTM Architecture

3.4 Cross Validation

We employed two different cross-validation methods to ensure the robustness and generalizability of our findings. The first method was KFold cross-validation, which involves randomly splitting the dataset into K (= 5) subsets while ensuring that the distribution of stress levels was balanced across each subset. The second cross-validation method was the Leave-One-Participant-Out (LOPO) approach, which involves training the model on data from all participants except one and evaluating its performance on the held-out subject. Given the participant's division into numerous groups and the presence of unique characteristics in each individual's response to the given stimuli, deploying a generalized model on unseen test subjects presented a significant challenge. To address this hurdle, we adopted a fine-tuning (ft) or calibration approach [4][17], incorporating a small portion of the unseen test data to train the model and tailor it to become more personalized or user-centric. This strategy allowed us to capture the peculiarities and individual traits of the test subjects, enhancing the model's ability to generalize and perform effectively on previously unseen data. The approach involves utilizing fractions of the unseen test data, specifically 5%, 10%, and 20%, to train the model in conjunction with the data from other participants. Conversely, for testing purposes, 95%, 90%, and 80% of the unseen participant's data are employed. For example, if p01 is the participant's data to be tested on, then 10% of its data is merged with the other participant's data as training data. The remaining 90% is used as testing data. This methodology allows for a controlled evaluation of the model's performance by systematically varying the proportions of training and test data from the unseen participants.

4 Results

The evaluation results of the two validation methods, KFold and LOPO with and without fine-tuning, using FCN, ResNet, Transformers, and LSTM models are presented in Table 1. It provides an overview of the performance comparison among the different models and the impact of incorporating the fine-tuning step to LOPO. The models were specifically trained to classify stress into three distinct classes: 'easy', 'medium', and 'hard'. Due to the differing durations of the relax session (90 seconds) and the arithmetic sessions (easy, medium, hard) (5 minutes each), there is a disparity in the time intervals, leading the models to become biased towards predicting lesser outcomes for the relax session. Moreover, several participants concluded their experiment with the relax session, and a considerable portion of them failed to stop the E4 recording following the session. Consequently, the relax session was not accurately labeled, leading to its exclusion during the model training phase.

Using the KFold cross-validation, we evaluated the performance of four different models, namely FCN, ResNet, Transformers, and LSTM, in predicting the three stress levels. The outcomes revealed accuracy rates of 85.05%, 95.05%, 59.12%, and 78.21%, respectively, indicating their varying levels of effectiveness. However, when utilizing the LOPO approach without fine-tuning, the FCN and ResNet models achieved lower accuracy rates of 41.16% and 35.52%, respectively, in classifying the stress levels. These results highlighted the limitations of the models when faced with the challenge of generalizing to unseen test subjects without the fine-tuning process. After incorporating the fine-tuning process to LOPO, notable improvements were observed. The FCN model's accuracy significantly increased to 47.63%, 53.81%, and 60.17% when fine-tuned with 5%, 10%, and 20% of the test data, respectively. Similarly, the ResNet model's accuracy soared from 35.5% to 60% after calibrating it with only 5% of the testing data from each participant. Furthermore, the accuracy improved to 80 and 85 percent when fine-tuning was performed with 10% and 20% of the data, as shown in the Table 1. These findings highlight the substantial impact of fine-tuning or calibrating on the performance of the models, resulting in more accurate and reliable stress classification outcomes.

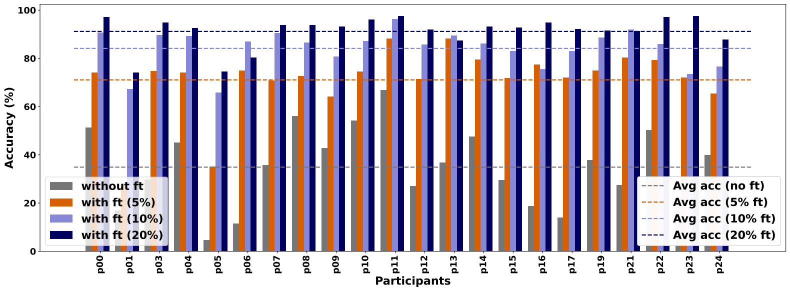


Figure 7: Results of LSTM using LOPO with and without fine-tuning (ft) for all participants.

Additionally, without fine-tuning, the LSTM model achieved an accuracy of 35.86% in classifying the stress levels. However, after fine-tuning the model with only 5% of the test data, the performance improved significantly to 71.09%. Eventually, as the fine-tuning percentage increased to 10% and 20%, the model's accuracy gradually increased to 84.12% and 91.17%, respectively. On the other hand, the Transformer model achieved an accuracy of 40.38% without any fine-tuning. However, the model did not experience a substantial change in performance with the application of the fine-tuning method.

Despite the initial accuracy achieved by Transformers, the fine-tuning process did not yield a notable improvement in its classification capabilities. The Figure 7 presents the distribution of predicted stress accuracies for each participant, showcasing the comparison between the model's performance with and without the fine-tuning process. By examining the bar chart, one can assess the impact of fine-tuning on the model's ability to accurately classify and distribute the stress instances in a generalized as well as a user-centric way. The confusion matrices for LSTM are displayed in Figure 8, which illustrate the changes in classification perfor-

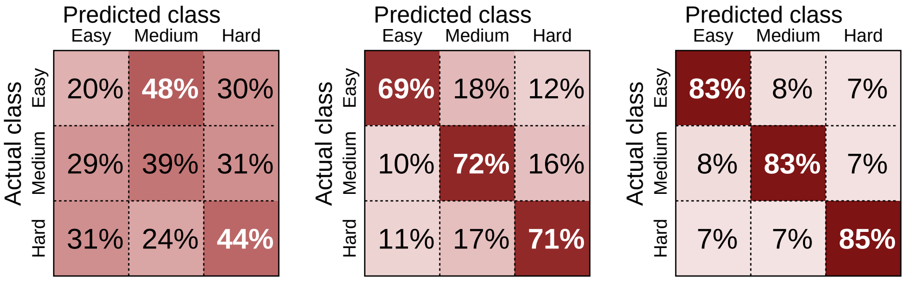
Table 1: Summary of classification results for three class stress detection using KFold and LOPO approach.

Model	**Validation method**	**Acc**	**F1**
FCN	KFold	85.05	0.848
	No ft	**41.16**	**0.483**
	5% ft	47.63	0.433
	10% ft	53.81	0.542
	20% ft	60.17	0.578
ResNet	KFold	**95.05**	**0.950**
	No ft	35.52	0.244
	5% ft	60.10	0.555
	10% ft	80.04	0.786
	20% ft	85.12	0.846
Transformers	KFold	59.12	0.572
	No ft	40.38	0.375
	5% ft	43.85	0.427
	10% ft	44.09	0.403
	20% ft	44.29	0.408
LSTM	KFold	78.21	0.781
	No ft	35.86	0.344
	5% ft	**71.09**	**0.709**
	10% ft	**84.12**	**0.839**
	20% ft	**91.17**	**0.911**

mance when fine-tuning is applied compared to when it is not. These matrices provide a visual representation of how well the model performs in accurately predicting the three distinct stress classes: easy, medium, and hard. By examining the confusion matrices in Figure 8, we can observe the distribution of predicted and actual stress classes and identify how the fine-tuning process enhanced the classification performance even with a small percentage of data.

5 Discussion

As part of our research, we investigated whether a student's physiological reactions differed when they faced varying levels of stress. We aimed to learn more about how stress affects student's well-being and performance. To substantiate these patterns, we designed an experiment to

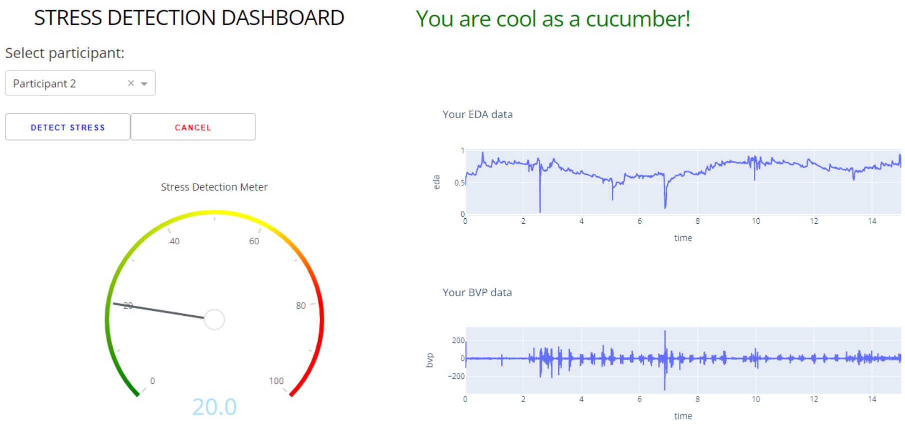


(a) No fine-tuning (b) With 5% fine-tuning (c) With 10% fine-tuning

Figure 8: Confusion matrix of LSTM (a) without fine-tuning (b) with 5% fine-tuning and (c) 10% fine-tuning.

analyze stress levels in students that involved a prominent stressor, mental arithmetic tasks. In addition to the regular arithmetic tasks, we used feedback, buzzers, and timers with clock-ticking sounds to induce extra stress in participants and to obtain distinct variations. The primary motivation behind designing this controlled system with specific ground truths was to eliminate reliance on participant's responses to questionnaires, as these can potentially impact the accuracy and integrity of the analysis. By employing a system with predefined ground truths, the study aimed to minimize potential biases and ensure a more objective assessment of the data. By implementing this multilevel approach, our models offer a more nuanced and comprehensive classification system, enabling a finer-grained understanding of different levels of stress experienced by individuals. This advancement allows for a more nuanced and detailed analysis of stress responses, leading to a richer understanding of the complex nature of stress and its varying intensities.

The deep learning models employed in this study have the advantage of eliminating the need for pre-processing or feature extraction from raw signals. However, accurately classifying the three stress levels from the time series data collected using the E4 wristband posed a significant challenge. The E4 wristband collects five different sensory data, and it was crucial to understand and effectively utilize this data to enable the models to distinguish between stress levels. One important consideration was the heart rate (HR) data, which can be derived from blood volume pulse (BVP) data sampled at 64Hz [29]. Given this relationship, we decided to exclude HR data from our analysis. To ensure compatibility with the Transformer and LSTM models, which expect inputs of multiple features with the same shape before passing through dense layers, the raw data provided to these models was downsampled to 4Hz. Unfortunately, downsampling resulted in a loss of data from the BVP and ACC sensors. On the other hand, the fully convolutional network (FCN) and ResNet implementations utilized all the raw data with their actual sampling frequencies. Despite these technical considerations, all the models demonstrated decent performance in accurately classifying the three stress levels, especially with fine-tuning.



STRESS DETECTION DASHBOARD

Select participant: Participant 2

You are cool as a cucumber!

(a) (b)

Figure 9: (a) Stress-meter (b) Customized message with EDA and BVP signals

This showcases the efficacy of the chosen approaches in leveraging deep learning models to tackle the complex task of stress level classification using the E4 wristband data.

The initial results obtained using a participant-independent approach did not yield satisfactory outcomes. This observation strongly suggests that the induced stress is highly dependent on the individual participants and their mental state during the study. To address these inter-participant variations, the models were calibrated using a small percentage of the test data through a fine-tuning process. Remarkably, this fine-tuning procedure had a substantial impact on the classification performance. By incorporating just 5% of person-specific data for training purposes, a significant improvement in performance was observed. This improvement underscores the importance of fine-tuning the models with unseen data to enhance their generalization capabilities. Furthermore, it highlights the participant-specific nature of stress, indicating that stress responses vary across individuals. Consequently, the calibration of the models using personalized data enables them to capture and accommodate these individual differences, leading to more accurate and reliable stress classification.

The experimental design consisted of four sessions, which included three stress levels and a 'relax' session. The purpose of incorporating the 'relax' session was to divide the participants into two groups: one group experiencing this session at the beginning of the experiment and the other group at the end. This approach aimed to investigate whether the timing of the 'relax' session influenced the levels of stress experienced by the two participant groups. Upon analyzing the results, no significant differences in stress levels between the two groups were observed. During the experiment, a notable number of participants completed their sessions with the 'relax' session. However, a considerable portion of these participants encountered difficulties in properly concluding the E4 recording following the conclusion of the 'relax' session. Consequently, inaccuracies arose in labeling the ‘relax’ session data, resulting in its exclusion from the model training phase.

In addition to implementing the models, we developed a prototype application using Python Dash to visualize and represent the stress prediction results obtained from each participant. The application interface, as illustrated in Figures 9a and 9b, was designed to provide a comprehensive overview of the stress levels over time. To assess and depict stress levels within the application, we opted for a distinct approach from the cross-validation techniques employed for evaluation in the study. We utilized a person-specific training method, wherein the data of the same participant was divided into training and testing sets using the KFold approach. By doing so, we ensured that the stress visualizing application could grasp and represent the individual differences and distinctiveness exhibited by each participant. The utilization of this person-specific training methodology aims to enhance the personalization and uniqueness of the application for every individual, as a general model may not be able to capture the individual differences.

Within the application, users had the flexibility to select participants from a list for analysis. Once a participant was chosen, the application displayed the dynamic changes in the stress meter and customized alert messages for each session, categorized as ‘easy’, ‘medium’, and ‘hard’. At the end of the process, the stress meter provided an overall indication of the participant’s stress level, while a feedback message was displayed on the screen to provide additional insights. The application also provided the option to view the EDA and BVP signals of the participants on the dashboard, allowing users to examine the physiological signals alongside the stress levels, providing a more comprehensive understanding of the relationship between stress and physiological responses. The development of this application prototype aimed to serve as a potential intervention mechanism for students in educational settings. Providing real-time insights into their mental state while performing various tasks, the students have an opportunity to understand, manage, and address their stress levels effectively. The power of this tool could be immense in aiding student’s well-being and enhancing their academic achievements by encouraging self-awareness and proactive stress control.

6 Conclusion

The presented paper focuses on a study that utilizes physiological signals obtained from an Empatica E4 wristband to analyze stress levels during mental arithmetic tasks. To predict these stress levels, end-to-end deep learning-based approaches were employed. The research involved a user study consisting of 25 university students, with the objective of inducing different stress levels categorized as ‘easy’, ‘medium’, and ‘hard’. Among the stress detection models proposed in the paper, both ResNet and LSTM exhibited remarkable predictive outcomes when utilizing the KFold, Leave-One-Participant-Out (LOPO) cross-validation technique and applying a fine-tuning or calibration approach with 5%, 10% and 20% of the test data to make the prediction more personalized.

Using KFold, ResNet and LSTM classified three classes with 95.05% and 78.21% accuracy, respectively. Additionally, ResNet achieved an accuracy rate of 85.12% and an F1-score of 0.846, while LSTM achieved an even higher accuracy rate of 91.17% and F1-score of 0.911 with 20% fine-tuning of the models. Across all four implemented models, utilizing fine-tuning with the LOPO cross-validation technique and employing 5%, 10%, and 20% of the test data consistently outperformed the baseline methods. Overall, the findings of the study demonstrate the effectiveness of the deep learning-based approach in predicting stress levels using physiological signals. The results highlight the superiority of the ResNet and LSTM models, along with the benefits of employing the fine-tuning or calibration technique in enhancing the accuracy of stress level classification in a generalized prediction approach. The paper's contributions shed light on the potential of deep learning models in stress analysis and provide valuable insights for further research in this domain.

Expanding the scope of this study to encompass a real-time stress detection system could have transformative implications in the realms of education and healthcare. Such a system has the potential to bring about substantial advancements in understanding and addressing stress-related issues. By integrating deep learning models into a real-time stress detection framework, the accuracy, and effectiveness of stress analysis could be greatly enhanced. To achieve this, it would be valuable to gather data from a diverse range of subjects across various stress-inducing scenarios. By continuously refining and expanding the dataset and incorporating advanced deep learning techniques, researchers can continuously enhance the accuracy and performance of stress detection models. The potential impact of a robust real-time stress detection system is far-reaching, with the potential to revolutionize how stress is understood, managed, and addressed in various domains, ultimately leading to improved well-being and quality of life.

References

[1] Ahn, J.W., Ku, Y., Kim, H.C.: A novel wearable eeg and ecg recording system for stress assessment. Sensors **19**(9), 1991 (2019)

[2] Ahuja, R., Banga, A.: Mental stress detection in university students using machine learning algorithms. Procedia Computer Science **152**, 349–353 (2019)

[3] Arsalan, A., Anwar, S.M., Majid, M.: Mental stress detection using data from wearable and non-wearable sensors: a review. arXiv preprint arXiv:2202.03033 (2022)

[4] Behinaein, B., Bhatti, A., Rodenburg, D., Hungler, P., Etemad, A.: A transformer architecture for stress detection from ecg. In: 2021 International Symposium on Wearable Computers, pp. 132–134 (2021)

[5] Bong, S.Z., Murugappan, M., Yaacob, S.: Analysis of electrocardiogram (ecg) signals for human emotional stress classification. In: Trends in Intelligent Robotics, Automation, and Manufacturing: First International Conference, IRAM 2012, Kuala Lumpur,

Malaysia, November 28-30, 2012. Proceedings, pp. 198–205. Springer (2012)

[6] Brantley, P.J., Waggoner, C.D., Jones, G.N., Rappaport, N.B.: A daily stress inventory: Development, reliability, and validity. Journal of behavioral medicine **10**, 61–73 (1987)

[7] Cho, H.M., Park, H., Dong, S.Y., Youn, I.: Ambulatory and laboratory stress detection based on raw electrocardiogram signals using a convolutional neural network. Sensors **19**(20), 4408 (2019)

[8] Cho, Y., Julier, S.J., Bianchi-Berthouze, N.: Instant stress: detection of perceived mental stress through smartphone photoplethysmography and thermal imaging. JMIR mental health **6**(4), e10140 (2019)

[9] Cohen, S., Kamarck, T., Mermelstein, R.: A global measure of perceived stress, journal of health and social behavior, vol. 24 (1983)

[10] Derogatis, L.R.: Brief symptom inventory: BSI. Pearson (1993)

[11] Dharmawan, Z., Rothkrantz, L.: Analysis of computer game player stress level using eeg data. In: 11th international conference on computer games: AI, animation, mobile, educational and serious games, La Rochelle, France, pp. 111–124. The University of Wolverhampton (2007)

[12] Duru, D.G., Duru, A.D., Barkana, D.E., Sanli, O., Ozkan, M.: Assessment of surgeon's stress level and alertness using eeg during laparoscopic simple nephrectomy. In: 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 452–455. IEEE, San Diego, California, USA (2013)

[13] Dzieżyc, M., Gjoreski, M., Kazienko, P., Saganowski, S., Gams, M.: Can we ditch feature engineering? end-to-end deep learning for affect recognition from physiological sensor data. Sensors **20**(22), 6535 (2020)

[14] Garcia-Ceja, E., Osmani, V., Mayora, O.: Automatic stress detection in working environments from smartphones' accelerometer data: a first step. IEEE journal of biomedical and health informatics **20**(4), 1053–1060 (2015)

[15] Ghaderi, A., Frounchi, J., Farnam, A.: Machine learning-based signal processing using physiological signals for stress detection. In: 2015 22nd Iranian Conference on Biomedical Engineering (ICBME), pp. 93–98. IEEE, Tehran, Iran (2015)

[16] Gjoreski, M., Gjoreski, H., Luštrek, M., Gams, M.: Continuous stress detection using a wrist device: in laboratory and real life. In: proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct, pp. 1185–1193. Germany, Heidelberg (2016)

[17] Gjoreski, M., Gjoreski, H., Lutrek, M., Gams, M.: Automatic detection of perceived stress in campus students using smartphones. In: 2015 International Conference on Intelligent Environments, pp. 132–135. IEEE, Prague, Czech Republic (2015)

[18] Gjoreski, M., Luštrek, M., Gams, M., Gjoreski, H.: Monitoring stress with a wrist device using context. Journal of biomedical informatics **73**, 159–170 (2017)

[19] Haak, M., Bos, S., Panic, S., Rothkrantz, L.J.: Detecting stress using eye blinks and brain activity from eeg signals. Proceeding of the 1st driver car interaction and interface (DCII 2008) pp. 35–60 (2009)

[20] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)

[21] Huang, M.X., Li, J., Ngai, G., Leong, H.V.: Stressclick: Sensing stress from gaze-click patterns. In: Proceedings of the 24th ACM international conference on Multimedia, pp. 1395–1404. Amsterdam The Netherlands (2016)

[22] Karthikeyan, P., Murugappan, M., Yaacob, S.: Ecg signals based mental stress assessment using wavelet transform. In: 2011 IEEE International Conference on Control System, Computing and Engineering, pp. 258–262. IEEE, Penang, Malaysia (2011)

[23] Karthikeyan, P., Murugappan, M., Yaacob, S.: Emg signal based human stress level classification using wavelet packet transform. In: Trends in Intelligent Robotics, Automation, and Manufacturing: First International Conference, IRAM 2012, Kuala Lumpur, Malaysia, November 28-30, 2012. Proceedings, pp. 236–243. Springer (2012)

[24] Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M.A., Kraaij, W.: The swell knowledge work dataset for stress and user modeling research. In: Proceedings of the 16th international conference on multimodal interaction, pp. 291–298. Istanbul, Turkey (2014)

[25] Kurniawan, H., Maslov, A.V., Pechenizkiy, M.: Stress detection from speech and galvanic skin response signals. In: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, pp. 209–214. IEEE, Porto, Portugal (2013)

[26] Linden, W.: What do arithmetic stress tests measure? protocol variations and cardiovascular responses. Psychophysiology **28**(1), 91–102 (1991)

[27] Lundberg, U., Kadefors, R., Melin, B., Palmerud, G., Hassmén, P., Engström, M., Elfsberg Dohns, I.: Psychophysiological stress and emg activity of the trapezius muscle. International journal of behavioral medicine **1**(4), 354–370 (1994)

[28] Mind Media: `https://www.mindmedia.com/en/solutions/research/blood-volume-pulse-ppg/`

[29] Peper, E., Harvey, R., Lin, I.M., Tylova, H., Moss, D.: Is there more to blood volume pulse than heart rate variability, respiratory sinus arrhythmia, and cardiorespiratory synchrony? Biofeedback **35**(2) (2007)

[30] Pickering, T.G., Devereux, R.B., James, G.D., Gerin, W., Landsbergis, P., Schnall, P.L., Schwartz, J.E.: Environmental influences on blood pressure and the role of job strain. Journal of hypertension.

Supplement: official journal of the International Society of Hypertension **14**(5), S179–85 (1996)

[31] Ramachandran, P., Zoph, B., Le, Q.V.: Swish: a self-gated activation function. arXiv preprint arXiv:1710.05941 **7**(1), 5 (2017)

[32] Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., Van Laerhoven, K.: Introducing wesad, a multimodal dataset for wearable stress and affect detection. In: Proceedings of the 20th ACM international conference on multimodal interaction, pp. 400–408. CO, Boulder, USA (2018)

[33] Secerbegovic, A., Ibric, S., Nisic, J., Suljanovic, N., Mujcic, A.: Mental workload vs. stress differentiation using single-channel eeg. In: CMBEBIH 2017: Proceedings of the International Conference on Medical and Biological Engineering 2017, pp. 511–515. Springer (2017)

[34] Shi, Y., Ruiz, N., Taib, R., Choi, E., Chen, F.: Galvanic skin response (gsr) as an index of cognitive load. In: CHI'07 extended abstracts on Human factors in computing systems, pp. 2651–2656 (2007)

[35] Sysoev, M., Kos, A., Pogačnik, M.: Noninvasive stress recognition considering the current activity. Personal and Ubiquitous Computing **19**, 1045–1052 (2015)

[36] Theodoros Ntakouris: `https://keras.io/examples/timeseries/` (2021)

[37] Ushiyama, K., Ogawa, T., Ishii, M., Ajisaka, R., Sugishita, Y., Ito, I.: Physiologic neuroendocrine arousal by mental arithmetic stress test in healthy subjects. The American journal of cardiology **67**(1), 101–103 (1991)

[38] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[39] Wang, Z., Yan, W., Oates, T.: Time series classification from scratch with deep neural networks: A strong baseline. In: 2017 International joint conference on neural networks (IJCNN), pp. 1578–1585. IEEE (2017)

[40] Wu, G., Liu, G., Hao, M.: The analysis of emotion recognition from gsr based on pso. In: 2010 International symposium on intelligence information processing and trusted computing, pp. 360–363. IEEE (2010)