User Stress Levels Detection by Physiological Sensing and Deep Learning

Master Thesis

Submitted by:

Rashmi Alur Ramachandra

Supervisor:

Prof. Dr. Prof. h.c. Andreas Dengel
Jayasankar Santhosh

RPTU Rheinland-Pfälzische
Technische Universität
Kaiserslautern
Landau

Master's Computer science
RHEINLAND-PFÄLZISCHE TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN-LANDAU
August 30, 2023

Declaration of Authorship

I, Rashmi Alur Ramachandra, declare that this thesis titled, “User Stress Levels Detection by Physiological Sensing and Deep Learning” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Rashmi Alur Ramachandra
Kaiserslautern, August 30, 2023

Acknowledgements

I would like to thank Prof. Dr. Prof. h.c. Andreas Dengel and Dr. Shoya Ishimaru for the opportunity to work on my thesis at Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Kaiserslautern. I am grateful to my supervisor, Jayasankar Santhosh, for his unwavering support and guidance throughout the completion of my thesis. His efforts were crucial to my smooth progress and the achievement of my goals. I cannot emphasize enough the critical role he played in my success.

Abstract

Stress is often associated with negative emotions and thoughts, which can affect the well-being of both the mind and body. It's a personal issue that affects many young people, particularly students, in today's society. To gain a better understanding of stress in educational settings, this study uses various methods to detect stress levels among students. Our goal is to gain a better understanding of the complex nature of stress in educational environments. We conducted an experiment with 25 participants, using an Empatica E4 wristband to record their physiological signals and determine their cognitive stress levels. 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 LOPO prediction accuracy, 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 for three-class classification. A prototype application has been developed that effectively displays dynamic stress fluctuations by utilizing the insights gained from prediction outcomes upon user-specific data. The application includes a stress meter, which enables users to visually understand their stress levels. It also delivers personalized alert messages to individuals based on their stress levels to ensure timely support and intervention.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Basic idea

Stress is often associated with negative thoughts and feelings and is considered a personal experience that can affect both mental and physical health. Psychological stress is a significant problem for young individuals, especially students, in today's digital world. The e-learning industry is rapidly growing over the few decades with continuous growth in the internet world. Nowadays, everything is available online in one touch, from a single-topic video to an entire semester. Especially, during the COVID-19 pandemic, e-learning has become crucial and necessary for students to continue and develop their knowledge. In today's digital age, students are required to attend numerous online classes and lectures, while also managing a heavy workload of assignments and challenging exams. Students can feel stressed due to the cumulative effect of their responsibilities, which can negatively affect their mental and physical well-being. It is imperative to recognize and address stress early on to ensure the welfare of the student.

On the other hand, over the past several years the use of wearable devices for collecting physiological parameters from humans are becoming customary.

Researchers have investigated how to use wearable devices to the fullest to facilitate numerous issues in different fields. These include basic physical activities [1], monitoring diseases [2], modeling behavior, service management and many more. Although wearable devices are very popular in these areas, research in the field of education [3] [4] or e-learning is just emerging.

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 [5] [6] [7]. These measures have proven to be valuable tools for evaluating the physiological responses associated with stress over the years [8]. 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. EEG, GSR and ECG data capture and interpretation technologies provide objective stress level assessments for researchers. This early stress identification enables timely interventions and effective support system implementation for students.

Modern education systems provide a wide range of customization possibilities but ignore the student's current cognitive load or mental state. Student's cognitive load greatly influences their ability to learn and successfully complete their subjects. In traditional learning methods, teachers detected cognitive load by interacting with students, asking questions to determine understanding levels and providing solutions if needed. As a result, students were more attentive in class, learning alongside peers of similar age. However, the impact of social distancing has made online learning difficult, leading to decreased engagement and potentially long-term academic repercussions. Additionally, online learning has made students more stressed, causing depression and effecting their well-being. The impact of this type of education has disturbed students on a more profound level.

As a traditional approach, machine learning has been used to learn the known data from sensors and use it to predict the information that will be available in the future. A number of developments have been made to the traditional machine learning approach with good results. This approach includes support vector machine (SVM), k nearest neighbor (KNN) algorithms and many more. Despite their advantages, they are not able to predict accurately the real-world dataset. To enhance their accuracy, a robust approach called deep learning or deep neural network is used in the field. Deep neural networks have demonstrated outstanding accuracy in forecasting the results for various datasets. Physiological measures such as heart rate, temperature and eyes contribute to predicting cognitive load.

1.2 Motivation

Isn't it wonderful to have a system that can detect our cognitive load by analyzing our physiological parameters and assist us in understanding our strengths and weaknesses? However, implementing a system that helps students identify their mental effort in the real-world is very challenging. Recent developments in the field of computer science and Artificial Intelligence provide a possible solution to these challenges. Binding up the advantages of wearable devices and the performance of artificial intelligence to implement such a system is the most effective possible solution.

Educators and institutions can achieve a deeper understanding of student stressors and customize educational environments by integrating physiological measures into stress detection methodologies. This knowledge can also help in creating personalized interventions and coping strategies to enhance students' mental health and academic progress 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 [9].

In our study, we employed the Empatica E4 wristband as a data collection 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'. To induce stress during the experiment, we utilized mental arithmetic tasks as stressors [10]. 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.

1.3 Contributions of the Thesis

The thesis aims to detect and understand student's stress levels in a variety of different scenarios. We aimed to detect variations across these stress levels. Our objective was to conduct a study to collect and analyze data from students in order to understand the relationship between different stress levels. We then used this data to create a predictive model that can identify and address sources of stress in students.

The key contributions of this study are:

- A meticulously crafted experimental protocol with a proven stressor and an array of tasks and stimuli that effectively induce varying levels of stress among participants. This guarantees consistent control and manipulation of stress levels throughout the entire experiment.

- Detailed analysis and comparison of several deep learning models for predicting stress levels, showing that these models can be trained effectively to predict stress levels accurately.

- Comparing a general classification scheme to a tailor-made approach that is fine-tuned according to unique user data.

- 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.

1.4 Structure of the thesis

Chapter 1 presents the basic idea, the motivation behind the research study and the contributions of the thesis. In Chapter 2, the fundamental background of stress and its correlation with human physiological signals are explained. Additionally, this section also includes some of the related work done in the history of stress detection research. Chapter 3 demonstrates the motivation for the user study, the experiments conducted for data collection and the workflow of the sessions involved in the experiment. Chapter 4 explains the approach used to analyze the data collected and highlights the model architectures along with the different cross-validation methods used in the study. Chapter 5 presents the classification of various deep learning models implemented to detect varying stress levels. Chapter 6 exhibits the end-user application developed using the predicted results. Chapters 7 and 8 discuss the findings and draws a conclusion on the experiment and results.

Chapter 2

Technical Background and Related work

Stress detection has a decade-long history in the field of computer science. Detection and analysis of stress can range from uncomfortable sensors to the most comfortable sensors and from uneasy situations to simple conditions and experiments. Detecting physical activities and diagnosing physiological features helps to understand human cognitive load. This theory can be used in understanding the mental efforts or even the emotional state of students while they are studying. Recent works have demonstrated a strong relationship between human physiological features and cognitive stress or emotion.

2.1 Stress measurement

Humans use their working memory to evaluate new information. Due to this evaluation, a burden is put on the working memory which is termed cognitive load/stress in many literatures. Cognitive load theory is a psychological term concerned with the use of a human's limited number of resources to acquire knowledge and skills required in various new situations. Hence, our working memory must process only the required information rather than distractions [11]. Approaches to measuring cognitive load can be divided into four categories: subjective measure, behavioral measure, performance measure and objective measure [3]. The subjective measure is the ability of the subject or participant to self-measure and report their stress level after performing tasks. Psychologists create these individualized questionnaires [12]. This report can be used as a ground truth to evaluate other cognitive loads. The behavioral measure is the activity of the user while performing a task, for example, a human tends to nod their head when listening to someone. Performance measures are based on the assumption that the subject's performance drops as memory capacity overloads. However, to remain successful, one must increase their efforts. Objective measurements include physiological signals from wearable and non-wearable sensors.

Stress detection is widely and traditionally done with subjective questionnaires. Psychologists have developed questionnaires for measuring different types of stress. Subjects are asked to complete a questionnaire to collect information. For the purpose of validating the objective measurements obtained from the sensors, researchers rely on psychological questionnaires. PSS (Perceived Stress Scale) [13] is one of the commonly used questionnaires to detect stress and overall emotional quotient [12]. Additionally, some of the other questionnaires developed by psychologists, according to [12] are Daily stress inventory (DSI) [14], Brief symptom inventory (BSI) [15], Acute stress disorder (ASD) [16] and Relative stress scale (RSS) [17]. However, using these questionnaires still has its limitations, as participants have to be knowledgeable enough to answer these questions honestly without any judgment or partiality. Also, the self-reporting method forces the subjects to pause their learning process or task and respond to survey queries. The behavioral measure tends to interfere with the primary task [18] and they may change depending upon the instructions given while performing tasks. In an ideal world, these limitations would make subjective and behavioral measurements less convincing without objective measurements. On the other hand, objective measurements or physiological sensors help to detect human stress or any emotion with wearable devices placed on the participant's body without any need for physical contact with them but they sometimes lack sensitivity.

2.2 Physiological data and Sensor-based devices

Electrodermal activity (EDA) also known as galvanic skin response (GSR) or skin conductance helps to read the variations in the human skin through the sweat glands. These variations can show a lot about the human's state of mind that reflects the sympathetic nervous system behavior [19]. The rise in EDA can be due to physical activities like running, sleeping and standing or emotional activities like excitement, stress, fear and anger. The human body data is measured by applying a very negligible and constant voltage to the skin that varies the conductance. Since the human body contains a large area of sweat glands, there are many possible locations to measure GSR like shoulders, wrists, fingers and feet. Examples of EDA devices that can be worn on the wrist are Empatica E4 and Fitbit (Figure 2.1a [20]). The two main characteristics of EDA signals are SCR (Skin Conductance Responses), a phasic component and SCL (Skin Conductance Levels), a tonic component. SCRs are the rapid and smooth events corresponding to either spontaneous skin conductance or an event-driven sympathetic reaction to some kind of stimulus. SCLs are the slow fluctuations in sweat glands due

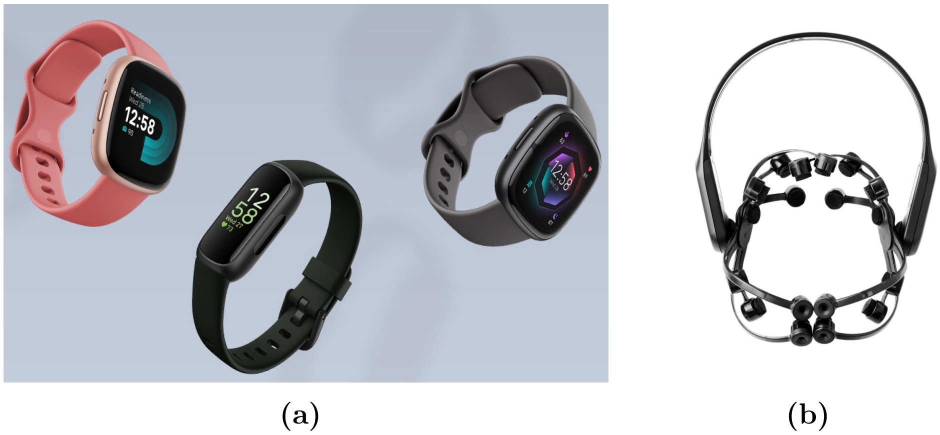


Figure 2.1: *(a) Fitbit (b) Emotiv-epoc*

to the interaction between tonic discharges caused by skin temperature and hydration. Many previous pieces of research have proved that the SCR and SCL consistently increase in stressful conditions.

Electroencephalogram (EEG) is a non-invasive method of recording brain electrical activities or waves. With the advancements in the medical field, wearable EEG devices are released on the market. These devices are easy to wear, head-mounted, comfortable and user-friendly. The dry, tiny and non-contact electrodes detect the electrical charges that occur due to the activities in the human brain. These wearables are both commercial or business-related and non-commercial. Commercial systems are headsets integrated with a varied number of electrodes used in research, for example, Emotivepoc (Figure 2.1b [21]), Mindo 4, Zeo headband and many more. Non-commercial system specifications are not available to all, e.g., Imec's headset and Neurokeeper. Researchers use EEG graphs to understand brain processes. These brain activities have a profound impact on human stress [22]. The Alpha and Theta bands from EEG signals are the prominent waves that show stress through changes between them [23].

Electrocardiography (ECG) is a non-invasive modality used to measure and monitor the functionalities of the heart. Similar to brain activities, the correlation between the human central system and the heart is used to measure human stress [24]. The three main sensors from a standard 12-lead electrocardiogram are placed at specific locations, right arm, left arm and left leg, on the human body to measure the heart's electrical activity recorded on electrodes. HRV (heart rate variability) is one of the significant parameters that help in detecting mental stress derived from ECG signals as it correlates with the human autonomic nervous system. A range of research has shown that apart from frequency-domain, the time-domain feature from ECG contributes more towards detecting stress. Some time-domain features extracted from ECG include heart rate, R peak amplitude and RR intervals. HRV is capable of monitoring any trivial change in human mental or physical conditions [25]. One of the most widely used methods for detecting stress is to measure the heart rate (HR). The RR interval (rhythm to rhythm) is the time elapsed between two consecutive R-waves of ECG signals and its reciprocal is heart rate. Heart rate is measured in beats per minute (bpm). A number of studies have shown a notable change in heart rate when under stress [26].

Photoplethysmography (PPG) is a technique to determine volumetric changes in blood vessels by transmitting infrared light through the tissue and determining the absorption of this light by the blood flowing through the vessels [27]. It is also used to measure the rate of blood flow, also referred to as BVP (Blood Vessel Pulse), as controlled by the heartbeat. The most common places to measure BVP are the fingers and earlobes. PPG devices are small in size, low cost and user-friendly. In a stressful situation, the human body releases a large amount of stress hormones that in turn increase human blood pressure.

Skin temperature is the measurement of heat on the outermost surface of the human body. Normal skin temperature ranges between 33 and 37 °C.

Skin being the largest organ of the human body plays a vital role in regulating temperature. Skin temperature changes are correlated with changes in heart activities, brain activities and sweat reactions. In turn, this correlation is related to human stress. Temperature can be measured at different locations on the surface of the skin like the forehead, arms, face and fingers. However, these locations give different results under stressful conditions. For example, the temperature measured on fingers and toes is usually lower than the temperature taken from the forehead or arms.

Other physiological signals that contribute to detecting stress in humans are EMG (Electromyography) [28], respiration and acceleration [29]. 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 has proven the relationship between EMG and human stress. Human respiratory rate, also called breathing rate, is the number of breaths humans take per minute. Unpleasant emotions alter breathing patterns and humans breathe faster under stressful conditions. 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. For stress detection, a gyroscope senses angular velocity along the x, y and z axes.

Tracking eye movement helps to measure where a person looks, which is called a point of gaze. These eye movements are converted to a stream of data that includes gaze point, gaze vector and pupil position. A reflection or vector is formed between the pupil and cornea due to the near-infrared light directed towards the pupil. Eye trackers can be screen-based or wearable glasses. In screen-based eye tracking, the participant sits in front of the monitor and interacts with it for eye movement measurement. Wearable glasses are similar to eyeglasses and can record eye activities by allowing the participant to walk around freely. Examples of eye trackers are Tobii Pro and EyeLink 1000.

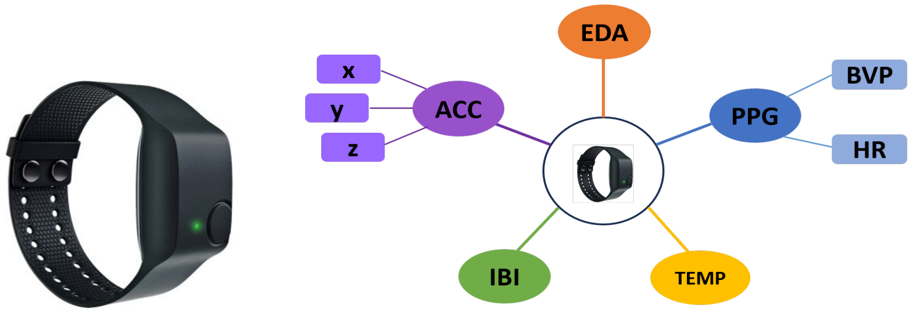
2.3 Biometric sensors

The biometric sensors used in our study are Empatica E4 wristband to measure the physiological variations in the participant's body and Tobii 4C with a Pro licence to measure the eye movements of the participants.

2.3.1 Empatica E4 wristband

The Empatica E4 (shown in Figure 2.2a [30]) is a wearable wristband that monitors physiological signals at a high level of accuracy and data quality in real-time with the help of sensors. It is comfortable to wear, weighs around 25g and captures data continuously. The E4 has an EDA sensor that reads the variations in human skin through sweat glands, a PPG sensor to measure BVP and HRV, an infrared thermopile to read the peripheral skin temperature and an accelerometer to acquire changes in movements in 3 axes (x, y and z) as shown in Figure 2.2b. It makes readability easier with an event mark button. This button helps to tag/log events correlated with the physiological signals. The raw physiological signals from E4 wrist band is shown in Figure 2.3

The web-based application from E4 helps to view the recorded data in graphs and download the raw data in CSV format. The tags.csv file records the event mark times or timestamps that are expressed in UTC. It contains one column which corresponds to the time when the button was pressed on E4. The data is synchronized with the session's initial time in the associated files. The EDA data is stored in EDA.csv and expressed as micro-Siemens (µS) at a sample rate of 4Hz. BVP data from PPG is stored in BVP.csv at



(a) (b)

Figure 2.2: *(a) Empatica E4 wristband (b) E4 sensors*

a sampling frequency of 64Hz. TEMP.csv contains data expressed in Celsius (°C) at a sample rate of 4Hz from the temperature sensor. HR.csv holds the average heart rate extracted from BVP signals. The folder also contains ACC.csv for 3-axis acceleration (sampling rate 32Hz) and IBI.csv files for inter-beat interval (no sample rate) data.

2.3.2 Tobii eye tracker

The eye tracker used in our experiment is Tobii 4C with a Pro license (shown in Figure 2.4 [31]). It is a small, portable and easy-to-use eye tracker with advanced features. It captures high-quality data with a 60Hz sampling frequency. The Tobii 4C-Pro is a screen-based eye tracker positioned on the computer screen. It measures eye movements while the participant interacts with the monitor. It can be connected to any laptop or desktop that supports USB 3.0 port. Ideally, it operates at a distance of 50 to 95 cm and is compatible with most of the eyewear like glasses and lenses. It has an additional feature that helps to capture the screen on any selected mouse click. The recorded data can be downloaded in CSV format. Some of the eye-tracking measurements include left-eye and right-eye coordinates, pupil diameter, blinks, etc.

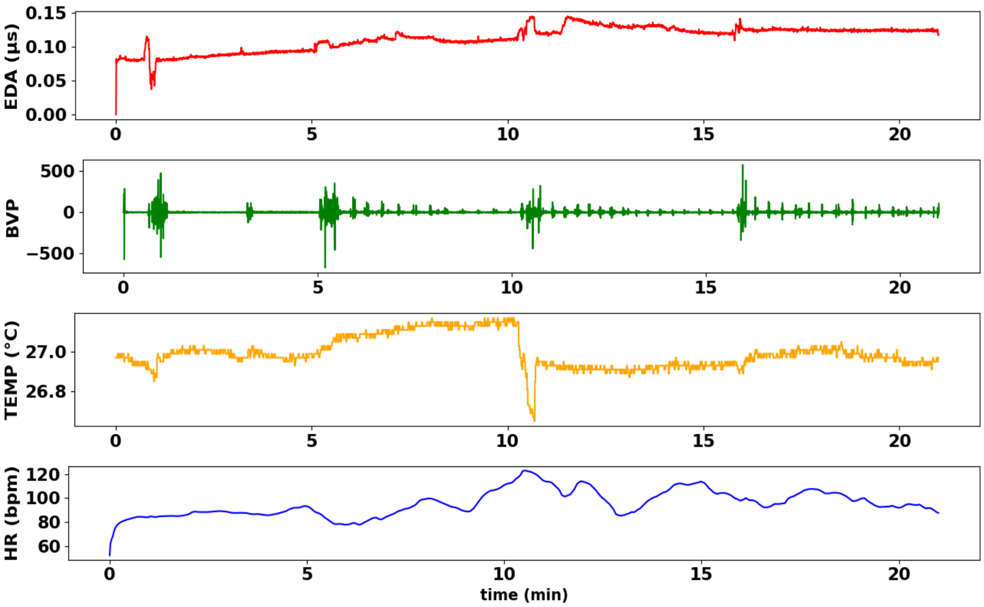


Figure 2.3: *The raw physiological signals from Empatica E4 wristband's EDA, BVP, TEMP and HR sensors.*

2.4 Stress detection using Physiological signals

A stress measurement system was developed by Hernandes et al. [32] in an office environment to determine the number of stressed and non-stressed calls performed by the employees using GSR. The traditional SVM model was used to classify the two classes and achieved an accuracy rate of 73.41%. A continuous real-time system was proposed by Panigrahy et al. [33] to monitor stress recorded in three different states – sitting, standing and sleeping using GSR. The system obtained 76.5% accuracy in classifying relaxed and stressed conditions. GSR along with other modalities can help to improve the accuracy, for example, Tang et al. [34] implemented a GSR system with two accelerometers achieving an accuracy of 94.7%, an improvement of +27.3% from using only GSR.

Hou et al. [35] proposed an innovative interface CogniMeter to monitor

Figure 2.4: *Tobii 4C Eye Tracker with a Pro license*

real-time human emotions, levels of mental workload and stress using only EEG signals captured from an Emotiv headset mounted on the user's head. The level of mental workload is paired with the recognition of positive or negative emotions to identify stress. An accuracy of 52.67% was achieved in classifying up to 8 emotions by combining fractal dimensions and statistical features from EEG signals. The five extracted features, absolute power, coherence, relative power, phase lag and amplitude asymmetry, from EEG cap signals gave an accuracy of 83.43% for multiple levels of stress detection using logistic regression, SVM and naïve bayers models [22].

A study by Karthikeyan et al. [26] achieved an accuracy of 96.41% for human stress recognition using ECG and discrete wavelet transform. The system classifies relaxed and stressed states using the k nearest neighbor classifier. This study uses heart rate variability as an extracted feature from recorded ECG signals. Also, the linear and non-linear HRV features from the time and frequency domain extracted from ECG signals were used to measure stress during oral examination by Castalso et al. [36]. The decision tree classifier outperformed other classifiers such as SVMs, Naive Bayes and multilayer perceptron by achieving an accuracy rate of 80%. ECG along with other modalities like GSR, facial expressions, BP, respiration and ACC enhances the experimental results. A real-time emotion recognition system using ECG and facial expression was developed by Tivatansakul et al. [37]. In this system, stress detection was used to determine the confusions in facial expression and effectively increase the preciseness of the system to provide relaxation service.

Acerbi et al. [38] conducted an experiment with 15 participants to detect stress during Maastricht Acute Stress Test using three commercial wearable sensors. Recorded heart rate variability signals are used to extract the mean of the heart rate feature. It was found that subjects with high-stress levels had significantly higher heart rates than those with low-stress levels. Similarly, a study by Schubert et al. [39] examines the effects of chronic and short-term stress using heart rate and HRV by comparing the time, frequency and phase domain complexities in 50 adults. The study concluded that the participant's heart rate significantly increased during the public speaking task as compared to the rest state. However, on the other hand, an experiment conducted by McDuff et al. [40] shows that heart rate does not play a significant role in detecting stress. The experiment included ball control and card sorting computer-based tasks to elicit stress which is recorded on contact-free cameras. Heart rate, heart rate variability and breathing rate was extracted from recorded PPG (photoplethysmography) signals. The study reported that there were substantial changes in heart rate variability during stressful conditions whereas there was no remarkable difference in the heart rates and breathing rates.

A multilevel mental stress detection using PPG was proposed by Zubair et al. [41]. The study used a Mental Arithmetic task to elicit stress in participants and achieved an accuracy rate of 94.33% in detecting five-stress levels using the traditional SVM model. A study by Steptoe et al. [42] proves the correlation between cardiovascular changes and mental stress levels. The researchers concluded that during the experiment, the group experienced an increase in blood pressure while performing stressful tasks. Additionally, a similar study by Hjortskov et al. [43] also concludes that there is a notable effect on blood pressure in participants performing computer-based tasks.

Vinkers et al. [44] proposed a study that shows the effect of core and peripheral body temperature on human stress. According to this study, under the same stressful conditions, the temperature of the fingertips decreased while the temperature of the upper arms increased. Hence, to consider skin temperature as a significant modality, it is important to measure them at the right location on the human body. A study by Karthikeyan et al. [45] identifies different levels of stress using changes in skin temperature. A probabilistic neural network achieved 88% accuracy in classifying four levels of stress conditions.

In a study by Sysoev et al. [46], behavioral and contextual data collected in real-life scenarios were used to determine stress non-invasively using smartphones. The data was collected from a 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. In a similar study proposed by Garcia-Ceja et al. [29], the system achieved 60% accuracy using a user-specific model relying solely on accelerometer data. Furthermore, Gjoreski et al. [47] mainly used accelerometers to classify not stressed, slightly stressed and stressed categories in the assessment of student stress. This three-level classification achieved 60% accuracy using 47 extracted features.

The behavior of the human eye is affected by different stressful situations. For example, in a study conducted by Haak et al. [9], in response to stressful emotions, eye blink rates and brain activity changed in the drivers. The pupillary response and the blink response are sensitive to the cognitive load. In general, humans tend to blink a lot when in stressful conditions. Also, the gaze fixation of the eye becomes unstable under the rhythm conditions [48]. Although the pupil diameter can dilate or constrict, the iris controls its movement of it. In the same way as physiological responses, pupil dilation is strongly related to cognitive and emotional arousal and cannot be controlled by the subject. A number of studies have demonstrated that pupil dilation is correlated with human affective states and used it as a marker to detect and assess stress [49]. In a study conducted by Tabbaa et al. [50] using the VREED dataset, the eye-tracking features outperformed physiological parameters like ECG in detecting a four-level classification problem for arousal and valence detection. In a study conducted by Torres-Salomao et al. [51], the participant's eye images were captured using a camera. It showed that pupils enlarged when solving difficult and stressful arithmetic tasks. Using both physiological and eye-tracking data helps improve classification results.

2.5 Stress Inducers or Stressors

Mental Arithmetic Task (MAT) is one of the most commonly used methods to induce stress in humans [52] [53]. Mental arithmetic tasks involve performing a series of basic arithmetic operations with increasing difficulty in order to increase the mental workload. This type of mechanism is simple and easy to implement. Montreal Imaging Stress Task (MIST), similar to MAT, developed by Dedovic et al. [54], was specially developed to induce psychological stress in a scanning environment. MIST is used as a stressor in many studies [55] [56].

Stroop Test is a neuropsychological test which has been developed by John Ridley Stroop in 1935 [57]. As a result of the Stroop effect, in psychology, the name of words interferes with the ability to identify the color of ink used to print the words by delaying the reaction time between automatic and controlled processing of information. Participants are tasked with naming the color of the word, not the word itself, as fast as possible. For example, if the word shown on screen or paper is “red” written in “green” ink, then the participant has to answer – green - but in reality, it is much easier to name the word that is spelled. Compared to reading, color naming is a powerful task as reading is an automatic process and color naming requires more cognitive effort. Various modifications have been made to the traditional Stroop task with the aim of understanding additional brain mechanisms as well as assisting in brain damage and psychopathology research.

The effects of music on human stress levels are numerous. Many researchers have demonstrated the positive and negative effects of music on human cognitive load. According to a study conducted in 1993 by Koelsch et al. [58], subjects were shown to have positive changes in stress hormones ACTH (Adrenocorticotropic) and cortisol levels when listening to music before and during stressful medical treatment. An investigation of the influence of music on biochemical parameters in humans was conducted in this study. A study conducted by Khalfa et al. [59] concluded that after a stress session, relaxing music is more effective in decreasing cortisol levels than silence. However, these researchers conclude the effect of one type of music - relaxing - on human stress. In reality, there are many genres like rock, rap, soothing, metal and classic. A recent journal by Asif et al. [60] investigated the effect of different genres of music, in English and Urdu, on human stress levels using EEG. The results show that English music has more influence in reducing stress as compared to Urdu music. Additionally, they also concluded that females’ stress levels were significantly lower than males, indicating that females are more sensitive to music than males. They reported an accuracy rate of 98.76% and 95.06% using logistic regression for two and three level classification respectively. Similar to music, video is one such parameter that elicits emotions in humans. Tabbaa et al. [50]

experimented on 34 participants using an immersive 360°video-based virtual environment (combination of audio and video) to elicit different emotions like joy, anger, relax and anxiousness. This dataset provided better results than other datasets that do not use VR effects.

The Trier Social Stress Test (TSST) is a stress-inducing laboratory procedure developed by Clemens Kirschbaum et al. [61] in 1993. This test includes two steps, in the first 5 minutes, called the anticipatory stress phase, participants have to prepare a presentation and in the next phase, participants have to present it and perform arithmetic tasks in front of the audience. This type of speech-based stressor is used in many studies to trigger emotions [62] [63]. Stress resulting from psycho-social situations such as public speaking, or facing people in a group, is a type of severe human stress. Many diseases include depression, cardiovascular disease and mood disorders as a result of exposure to social stressors [64]. These type of stressors affects the human autonomic nervous system and neuroendocrine system activating a fight or flight response in humans [65].

Similarly, pictures are one of the stressors that can elicit emotions in humans. One example of such a type of stressor system is the International Affective Picture System (IAPS) developed by the National Institute of Mental Health Centre for Emotion and Attention. It is a database of 956 color pictures ranging from everyday objects to portraits - plants to landscapes with a 9-scale rating for arousal, valence and dominance. Participants taking part in the experiment must rate how calm or excited, how happy or sad and how pleasant or unpleasant each picture they see. In general, the Self-Assessment Manikin (SAM) [66] rating scale is used as the rating procedure. For stress recognition experiments, IAPS has proven very effective for inducing stress [67] [68]. Likewise, Wisconsin Card Sorting Test (WCST) is also a popular neuropsychological test that measures the cognitive load on working memory [69]. This test allows the researchers to understand and detect cognitive functions of the brain such as attention, memory and visual processing.

Furthermore, there are many other real-time activities that are used to detect stress like car driving, running, sitting and standing [70]. Stress measurement in the students can be mostly elicited with university examinations like oral exams, puzzles, computer games, report writing and preparing presentations. Similarly, there are many studies that induce stress in employees while performing jobs in call centers, checking email and information search.

2.6 Publicly available datasets for Stress Detection

2.6.1 WESAD

Currently, the research community has accumulated a small number of datasets for human stress assessment that are publicly available. Among them, WESAD (WEarable Stress and Affect Detection) [9] stands out. WESAD contains data from 15 participants for wearable stress and effects detection. This dataset contains physiological and motion features collected from wrist and chest-worn devices acquired in a laboratory setting. The physiological features include ECG, EDA, EMG, respiration, BVP, skin temperature and the motion features include 3-axis acceleration. Additionally, the study also consists of self-reported data from all subjects. The data were recorded in three different conditions, 1) baseline condition: reading magazines to induce a neutral affective state, 2) amusement condition: watching eleven funny clips lasting for 392 seconds, 3) stress condition: a Trier Social Stress Test (TSST) which contains public speaking and mental arithmetic tasks.

2.6.2 SWELL-KW

SWELL-KW (Smart reasoning system for WELL-being at work and at home Knowledge Work) [71] is a multi-modal dataset for stress and user modeling. SWELL contains data from 25 subjects performing tasks like writing reports, reading emails, searching for information and preparing presentations under time pressure and interruptions. The features extracted to detect stress are facial expressions from cameras, HRV, GSR, body postures from a Kinect 3D sensor and computer logs. The dataset contains raw data and pre-processed data with extracted features. Also, subjective experience with validated questionnaires is used as ground truth. A total of three hours of data was recorded for each participant, subdivided into three one-hour blocks. During each block, the participant was given a relaxing eight-minute period of time to relax before beginning his/her work. During each block of the experiment, participants wrote two reports and presented one presentation.

2.6.3 CLAS

Another publicly available dataset, CLAS (Cognitive, Load, Affect and Stress) by Markova et al. [72], was developed to assess certain states and conditions of the human brain. The study contains data from 62 healthy participants performing three interactive tasks and two perceptive tasks. As part of the interactive tasks, people are asked to solve mathematical problems, logical problems and a Stroop test that induces a certain amount of cognitive load in their brains. Perceptive tasks include images and audio-video catalysts to elicit emotions in four-class quadrants of arousal and valence [73]. The dataset contains physiological signals like ECG, EDA, PPG and accelerator data. The study aims to integrate functionalities that allow to detection of stress conditions and human emotions automatically. It will also assess the degree of concentration of humans, their cognitive load through a variety of logical and mathematical problems under time limits.

2.7 Stress detection using deep learning

Machine learning has traditionally been used to predict future information based on known data from sensors. While several developments have improved the traditional approaches, they still fall short of accurately predicting real-world datasets. To enhance accuracy, deep learning or deep neural networks are now being utilized. These networks have proven to be highly accurate in forecasting results for various datasets, particularly when considering physiological measures such as heart rate, temperature and eye movement to predict cognitive load. Deep learning networks are able to quickly analyze and process large amounts of data, identify patterns and make more accurate predictions. Additionally, these networks are able to learn from their mistakes and adjust their predictions accordingly. Lastly, deep learning networks are capable of recognizing more nuanced patterns within the data, leading to more accurate predictions.

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. [74], 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 preprocessing or feature engineering involved. Similarly, in a paper by Dziezyc et al. [75], 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. [76] presented an end-to-end deep learning model for emotion recognition based on two publicly available datasets, WESAD [9] and SWELL-KW [71]. The research includes developing a neural network based on convolutional layers and multi-head transformers that are applied to ECG signals.

Chapter 3

User Study and Ground truth Collection

3.1 Motivation

At the start of our study, we asked the question: what is the most effective and meaningful way to study the cognitive stress levels of students? According to Karthikeyan et al. [26] [45], one of the simplest ways to induce stress is by asking participants to solve arithmetic tasks under time pressure with feedback popping up on the screen. This type of stressor can be applied to any group of participants, but it is particularly relevant to the target group in our study, students. We, therefore, decided to use this method as our stressor in our study. We asked participants to solve arithmetic tasks of increasing complexity under time pressure, with their results being displayed on the screen. This allowed us to measure the cognitive stress effects on our participants. Our study hypothesizes that cognitive load/stress increases with increasing difficulty in task levels and time restraints in the task. The reasoning behind this choice of stressor is that cognitive tasks can be easily quantified and measured, allowing us to objectively observe stress effects on our participants. Additionally, the increasing task levels allowed us to measure how cognitive stress increases with higher difficulty and compare the results to our hypothesis. In order to put this thought into action, we designed an experimental interface that induces varying stress levels using mental arithmetic tasks along with feedback and buzzers popping up on the screen.

3.2 Participants

To collect data on stress detection, we conducted a psychological experiment inducing different levels of stress in participants. Initially, an invite and a Google form were sent to all university students via different mailing platforms. As the experiment was simple, there were no participation criteria. Students who were interested in taking part selected a time slot. 25 participants - 14 male and 11 female - all aged between 21 and 31 years, volunteered to participate in the experiment. Most of our participants were university students majoring in computer science. All participants were given an overview of the study along with a consent form prior to the experiment. Participants who agreed and signed the consent form were allowed to proceed with the experiment.

3.2.1 Apparatus and Setup

The experiment was conducted in a dedicated room with all the apparatus - table, chair, monitor, keyboard, mouse, earphones, eye tracker and E4 wrist band as shown in Figure 3.1.

Empatica E4:

The data stored in E4's internal memory can be synced with the cloud

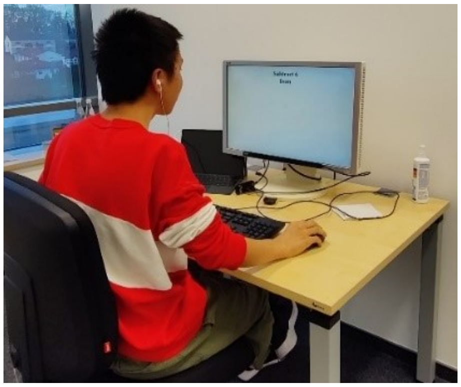


Figure 3.1: *Experimental setup*

using the Empatica Manager application via USB. This stored data can be accessed from its cloud platform called Empatica Connect. Empatica Connect helps to view the recorded data in graphs and download the raw data in csv format along with timestamps. This web-based application can be accessed from anywhere. The data from EDA, BVP, TEMP and ACC were used for implementation.

Eye Tracker:

Tobii 4C-Pro is compact and one of the easiest ways to collect eye movement data. It allows us to include different modes like connecting the eye tracker, getting eye tracker status, starting recording, getting recording status, capturing the screen, stopping recording and disconnecting the eye tracker.

3.2.2 Experimental Procedure

The participants visited the experiment room in the time slot they selected. The experimental procedure and all the apparatus were explained verbally. Before starting the experiment, individual eye gaze calibration was done where they were instructed to look and follow the dots appearing on the

Table 3.1: *Division of 25 participants*

Groups	**Participants**	**First main session**	**Second main session**	**Stress levels**
1	p00,p01,p02,p03	relax	stress	easy-medium-hard
2	p04,p05,p06,p07	relax	stress	hard-medium-easy
3	p08,p09,p10,p11	stress	relax	easy-medium-hard
4	p12,p13,p14,p15	stress	relax	hard-medium-easy
5	p16,p17,p18,p19,p24	relax	stress	medium-easy-hard
6	p20,p21,p22,p23	stress	relax	medium-easy-hard

screen. We asked the participants to keep their eyes on the screen throughout the experiment to ensure that the eye tracker collects as much data as possible accurately. After calibration, we asked the participants to wear the Empatica E4 wristband on their non-dominant hands. 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. After that, participants had to be engaged with the tasks coming one after the other for 20 to 25 min. The 25 participants were mainly divided into 2 groups depending upon the order of relax and stress sessions; 1) relax-stress, 2) stress-relax. These two groups were further divided into 3 groups depending on the stress levels; 1) easy-medium-hard, 2) medium-easy-hard, 3) hard-medium-easy. The participants started the experiment depending on their group as shown in Table 3.1. At the beginning and end of the relax and stress session, the participants were asked to tag/press the button on the E4 wristband. Each press is indicated by a red blink. At the start of each session, 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.

In between each session, participants took a few-second break. Along with the alert messages, the buzzer and ticking sounds were used to induce extra stress or anxiety in the participants. The pre- and post-questionnaire forms

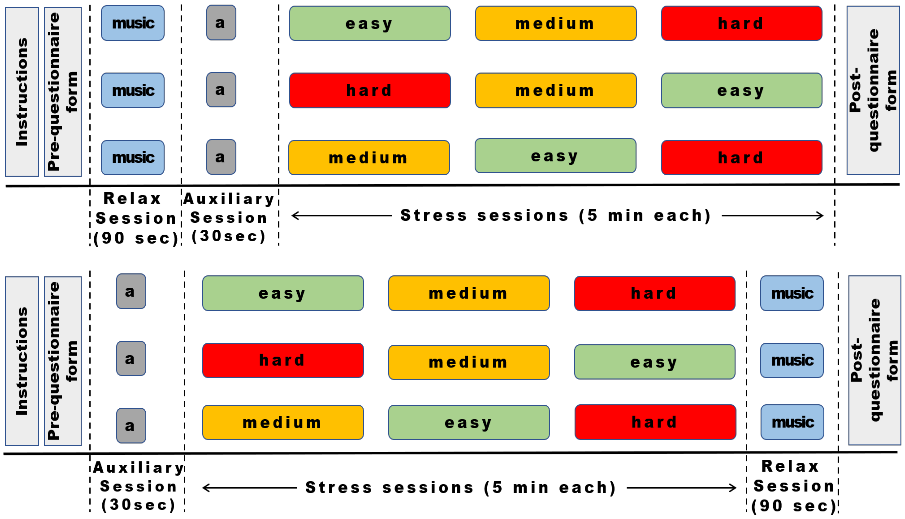


Figure 3.2: *Workflow of the application design*

had a Likert scale from 1 to 7 (where 1 is the least and 7 is the highest) and a drop-down menu to select their option. As a reward for their participation, each student received a 10-euro Amazon voucher.

3.3 Workflow

Figure 3.2 shows the workflow of the designed application. The application had a series of sessions including relax, stress and pre-post questionnaire forms. Each session was timed and detailed information regarding the timing of stress sessions is illustrated in Table 3.2.

3.3.1 Sessions

1. Pre-questionnaire form: All participants were asked to fill few personal questions like name, age, sex and rate a few general questions like “Do you listen to music?”, “How good are you at basic math?”, “how

Table 3.2: *Details of Stress Levels (easy, medium and hard)*

Tasks	**Total no.s in series**	**Each no. range**	**no. to add/sub/ mul**	**Each no. shown for**	**Time given to input series**
Addition	3	1 to 10	less than 5	3.5 sec	16.5 sec
Subtraction	4	1 to 20	less than 8	3 sec	15 sec
Multiplication	5	1 to 13	2 to 5	3 sec	15 sec

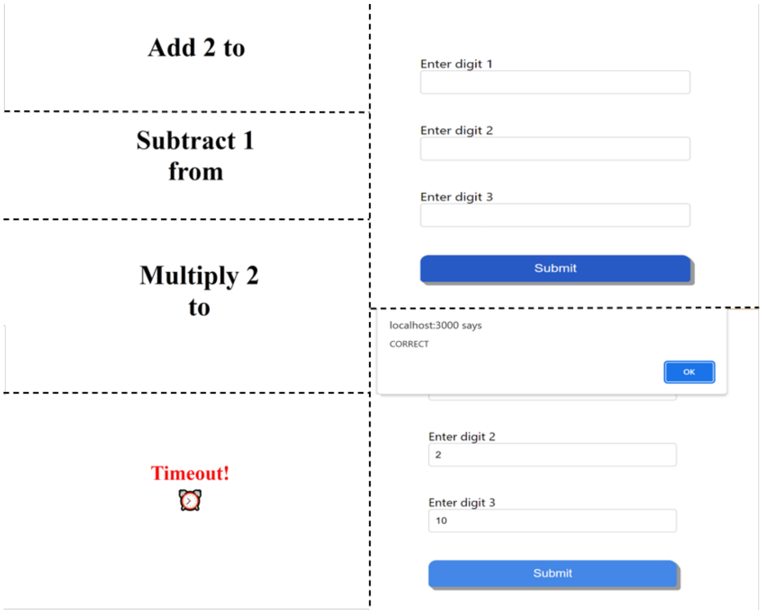
mentally alert do you feel at the present time?" and "Do repetitive soothing tracks make you feel asleep?"

2. 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.

3. Stress: The stress session included 3 levels of arithmetic tasks. All levels were 5 min 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 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 3.3.

   (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



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 3.3: *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)*

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.

4. 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 complimentary session that would assist the participants to concentrate

on the upcoming arithmetic tasks, we did not use an eye tracker or wristband. This session lasted for 30 seconds.

5. Post-questionnaire form: To gather feedback from participants about the experiment, we asked them to fill out a questionnaire after completing all the tasks. A few of the questions asked in this session were "Which gaming session was easier? (add/sub/mul)", "Which gaming session was more difficult? (add/sub/mul)", "Did the time limit make you feel anxious?"

Chapter 4

Methodology

The goal of the thesis is to identify and comprehend the levels of stress experienced by students in various situations. Our objective is to uncover differences in these stress levels using different deep-learning models. In this chapter, we will discuss the methodology used to pre-process the raw data (section 4.1). In section 4.2, we will discuss the architectures of FCN, ResNet, Transformers and LSTM. In section 4.3, all the cross-validation methods used in the study are explained.

4.1 Data pre-processing

The data collected from the Empatica E4 wristband was pre-processed and segmented using a sliding window of the length of 30 seconds without overlap. Our methodology utilizes an end-to-end deep learning approach that allows us to derive valuable and useful information directly from the raw signals without slogging through the hefty process that includes the manual extraction of features from the data. An end-to-end deep learning approach helps process the inherent insights contained in raw sensor data by omitting manual feature engineering steps. This process facilitates the deep

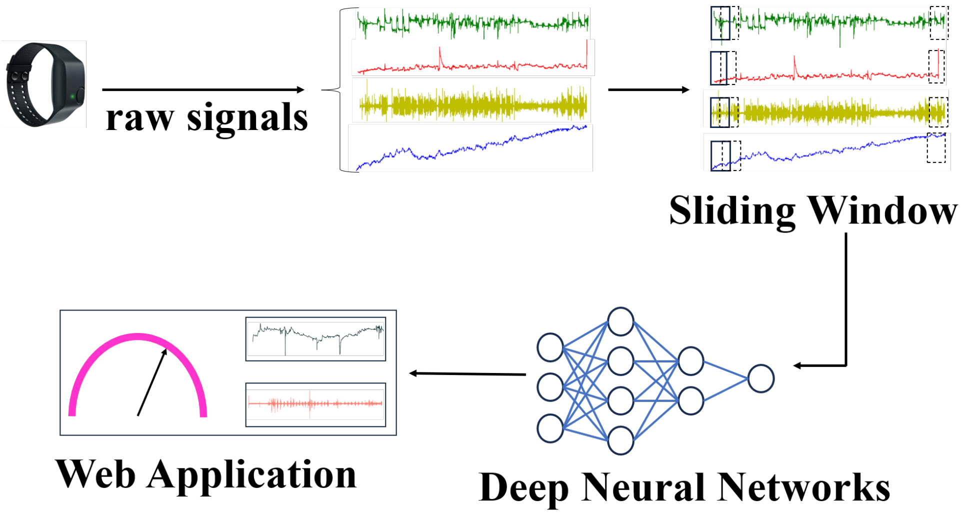


Figure 4.1: *Pipeline*

learning architecture's ability to extract and learn significant information independently, thus reducing complexity and eliminating potential biases introduced by human-designed features, allowing raw data to be analyzed more comprehensively and objectively. Additionally, deep neural networks can extract meaningful features that human experts sometimes cannot derive in a small amount of time accurately. The pipeline is shown in Figure 4.1. We collected the physiological signals like EDA, BVP, TEMP, HR and 3-axis ACC data from Empatica E4. Figure 4.2 illustrates the variations in the raw sensor signals during different stress sessions in the study. The eye tracker data was not included in the analysis as participants primarily focused on the keyboard rather than on the screen, resulting in mostly null raw eye data.

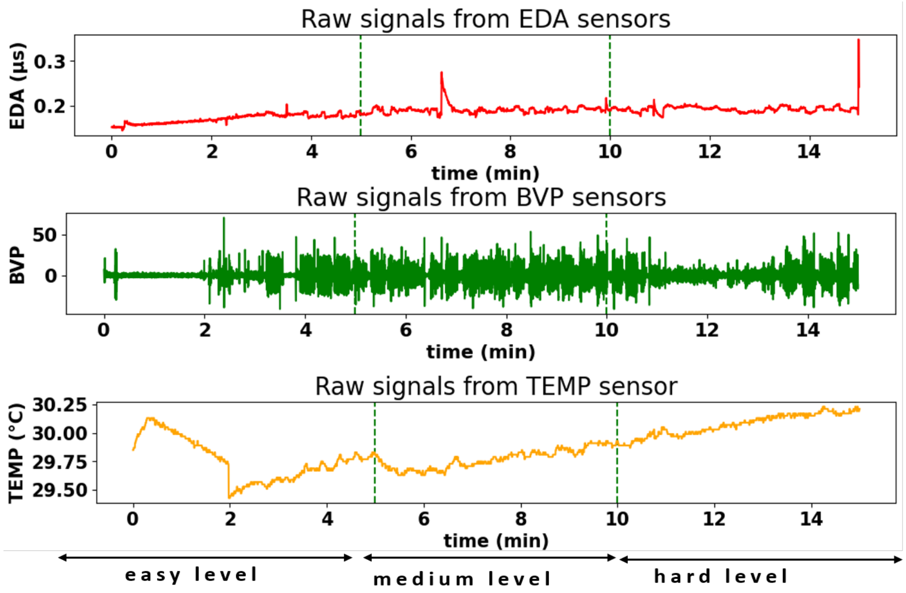


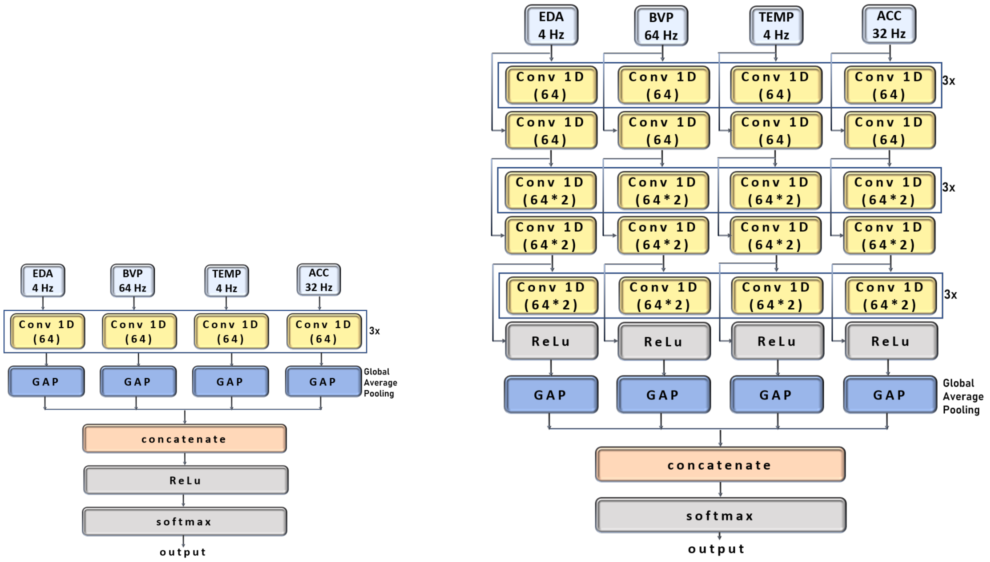
Figure 4.2: *The raw physiological signals from EDA, BVP and TEMP sensors separated by different sessions of stress for a single participant (p00).*

4.2 Model Architecture

Through their hidden layers, deep neural networks acquire high-level features without the need to engineer them. Therefore, the workflow becomes simpler and less manual effort is required, while the probability of capturing relevant information increases. The collected data was analyzed using various models, including ResNet [75] [74], FCN [75] [74], Transformers [77] [78] and Long Short-Term Memory (LSTM) [79].

FCN

Figure 4.3a shows the architecture of the FCN model that we implemented. The sensory data from EDA, BVP, TEMP and ACC were taken at their standard sampling frequency for the FCN. The FCN model consists of three



(a) **(b)**

Figure 4.3: *(a) FCN Architecture (b) ResNet Architecture*

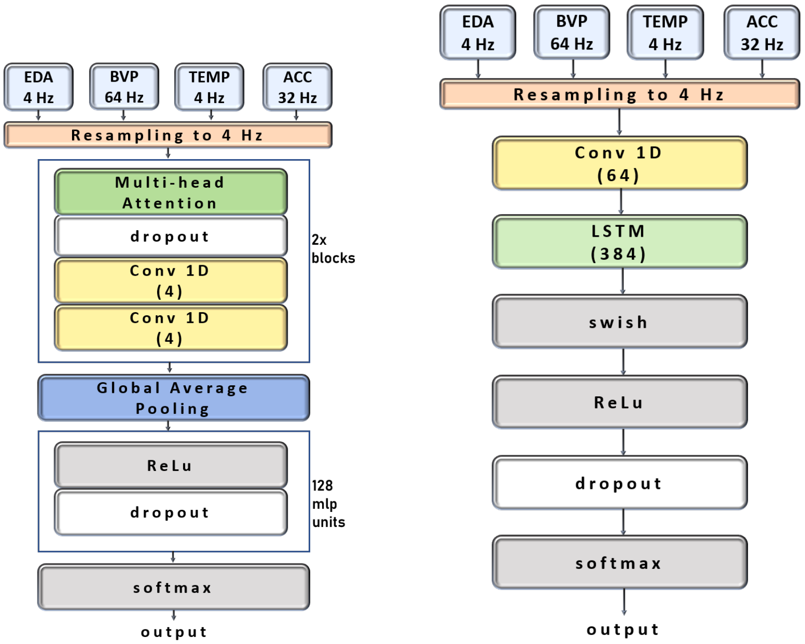
convolutional blocks for each signal of kernel size 3 and filter size 64. This is followed by a global average pooling layer. Since the signal contains multiple features, the four Global Average Pooling (GAP) branch layers are concatenated and fed to one or more fully connected dense layers. Each convolutional layer applies a set of filters to the input time series. The output of each layer is passed through a non-linear activation function such as the Rectified Linear Unit (ReLU) function. The convolutional layers output is processed by the fully connected layers to complete the final classification.

ResNet

The diagram in Figure 4.3b shows the architecture of ResNet. The network is highly dense, with multiple residual blocks and shortcut connections within each block. Our model contains three residual blocks where each block contains three successive convolutional layers with 64, 128 and 128 filters, respectively. The kernel size of convolutional layers also varies for each block, 8, 5 and 3 for each layer. Each convolutional layer is normalized batch-wise with a ReLU activation function. The three successive convolutional layers and an additional convolutional layer attached to the three successive convolutional layers receive the same input. A global average pooling layer is incorporated to down-sample the feature maps and capture the most relevant information. These layers are implemented for all the features, ACC, BVP, EDA and TEMP and concatenated just after the global average pooling layer. Towards the end of the architecture, fully connected layers are used to map the learned features to the desired output. The output layer provides the final prediction based on the processed physiological data.

Transformers

To process the raw input signals from EDA, BVP, TEMP and ACC in Transformers, the signals were downsampled to 4Hz. These signals were then passed through two Transformer blocks. The transformer block consists of a multi-head attention layer, dropout layers, a normalization layer and two 1D convolutional layers. Initially, the attention mechanism was used in Natural Language Processing (NLP) to translate text or in a chatbot to answer inputs. A promising result has emerged from its application to time series prediction. The multi-head attention layer helps the model share the input sequence across multiple vectors without relying completely on a fixed-length vector. A kernel with a filter size of 4 is used in the convolutional layers. The transformer block is coupled with a pooling layer and an MLP block. At last, the signals are passed through a Softmax layer as shown in Figure 4.4a.

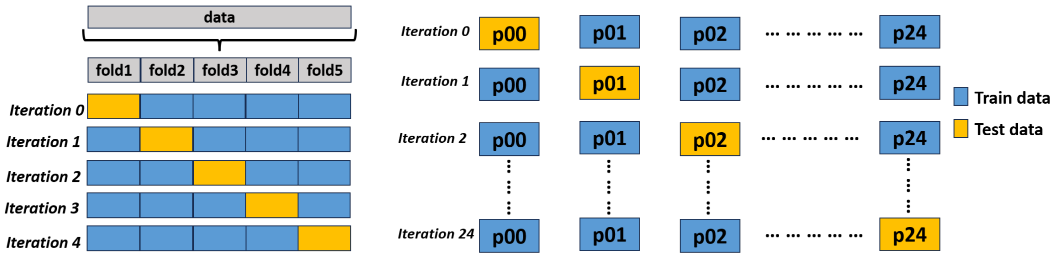


(a) (b)

Figure 4.4: *(a) Transformers Architecture (b) LSTM Architecture*

LSTM

Figure 4.4b illustrates the architecture of an LSTM model. Similar to Transformers, the inputs for LSTM 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 [80] and ReLu. There are 384 units in the LSTM block. The swish activation function helps deep learning models accurately perform predictions because of its smooth and non-monotonic behavior. It is represented in the form of f(x) = x · sigmoid(x). Subsequently, the output is passed through a dropout layer and dense layer for further processing.



(a) *KFold* (b) *LOPO*

Figure 4.5: *Cross-validation techniques utilized (a) KFold and (b) Leave-One-Partcipant-Out*

4.3 Evaluation Protocols

We employed two different types of cross-validation methods to ensure the robustness and generalizability of our findings. The first cross-validation approach is participant-independent and the other cross-validation method is participant-dependent. The participant-independent cross-validation approach ensures that the model is generalizable across different individuals and datasets. It also ensures that the results are applicable to a wider population as they are less affected by individual differences. On the other hand, the participant-dependent approach assures that the model is robust to variability among a group of participants. It allows us to take into account individual differences and better understand the effects of our experiment. The participant-independent approach can be used to determine the average effect of a given experiment, while the participant-dependent approach can be used to understand individual variance in response to a given experiment. Additionally, the participant-dependent approach is carried out in two ways, a) fine-tuning or calibrating each participant's data and b) the KFold method. The fine-tuning approach is useful to capture individual differences in response, whereas the KFold method allows for a more comprehensive evaluation of the data set.

Participant-independent approach

The participant-independent approach includes Leave-One-Participant-Out (LOPO) cross-validation, which involves training the model on data from all participants except one and evaluating its performance on the held-out subject as shown in Figure 4.5b. This procedure is done for all 22 participants and the results are averaged. All 22 participants had gone through the same arithmetic tasks but in a different order. It is interesting to note that although the tasks were the same for some groups, there is no high correlation between their performance and physiological signals. Given the participant's division into numerous groups and the presence of distinct characteristics in each individual's response to the given stimuli, deploying a generalized model on unseen test subjects presented a significant challenge. This has been a prominent issue for the past several years. In a paper by Hernandez et al. [32], a similar disassociation of stress levels was found among multiple call center employees. All participants reported stress levels differently, despite having the same job profile. There were similarities between participants in day-to-day work, but huge differences between them. This issue was analyzed using two methods: 1) modifying the loss function of the model and 2) emphasizing training samples from more similar individuals.

Participant-dependent approach

The participant-dependent approach is more about using a particular participant's data for analyzing his/her personalized affective states. This approach emphasizes the importance of individualized data for each participant. It focuses on the individual participant or user and his/her response to certain stimuli. It involves the use of sensors and other biometric measures to understand the user's emotions and reactions. This method helps to develop personalized services that are tailored to the user's needs.

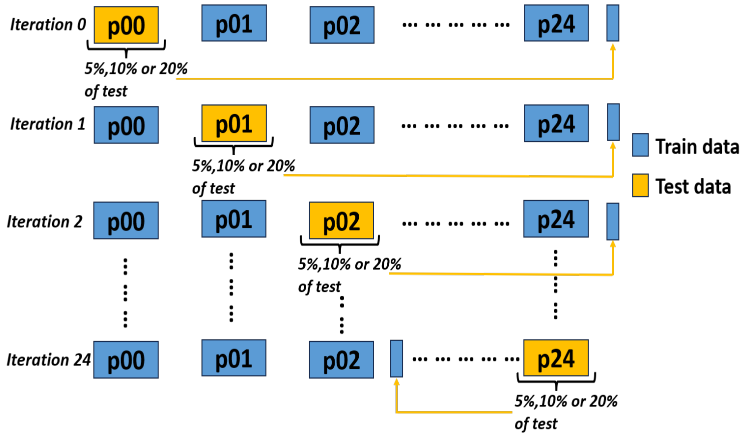


Figure 4.6: *Fine-tuning or calibration technique on Leave-One-Partcipant-Out*

a) Fine-tuning or calibration technique: To address such an imbalanced hurdle caused by LOPO, we adopted a fine-tuning or calibration approach [47] [76] as an extension to LOPO. This method incorporates 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 as shown in Figure 4.6. 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 5% of its data is merged with the other participant's data as training data. The remaining 95% 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.

b) KFold: The second participant-dependent method used is KFold cross-validation, which involves randomly splitting the dataset into K (= 5) subsets while ensuring that the distribution of stress levels is balanced across each subset as shown in Figure 4.5a. KFold cross-validation is an efficient way to measure the accuracy of model performance, as the process of randomly splitting the dataset into multiple subsets helps to minimize potential bias. It allows for an unbiased assessment of model performance, as each subset of data can be tested separately and the results combined to provide an accurate overall result. This method ensures that the model is trained and tested on different data, providing the most accurate evaluation of performance.

Chapter 5

Results and Analysis

The evaluation results of the LOPO, both with and without fine-tuning and the KFold approach using FCN, ResNet, Transformers and LSTM models are presented in Table 5.1. It provides an overview of the performance comparison between the different models and cross-validation methods. Additionally, Figure 5.1 demonstrates the prediction accuracy of FCN, ResNet, Transformers and LSTM using KFold cross-validation on each fold. The models were specifically trained to classify stress into three distinct classes: ‘easy’, ‘medium’ and ‘hard’ using physiological data obtained from the E4 wristband. As the ‘relax’ session lasted for 90 sec and arithmetic sessions (easy, medium, hard) lasted for 5 min each, there is a variation in the time stamps and the models tend to bias to predict lesser results on the relax label. Additionally, in the ‘relax’ session no mental arithmetic tasks were used. It was only designed to distribute participants into as many groups as possible. Therefore, we have omitted the ‘relax’ session from training the model. It was necessary for the participants to input their answers from the keyboard while performing mental arithmetic tasks. Many participants, however, were eager to enter the answer within the given time, so their eyes were more on the keyboard than their computer screen. Therefore, most

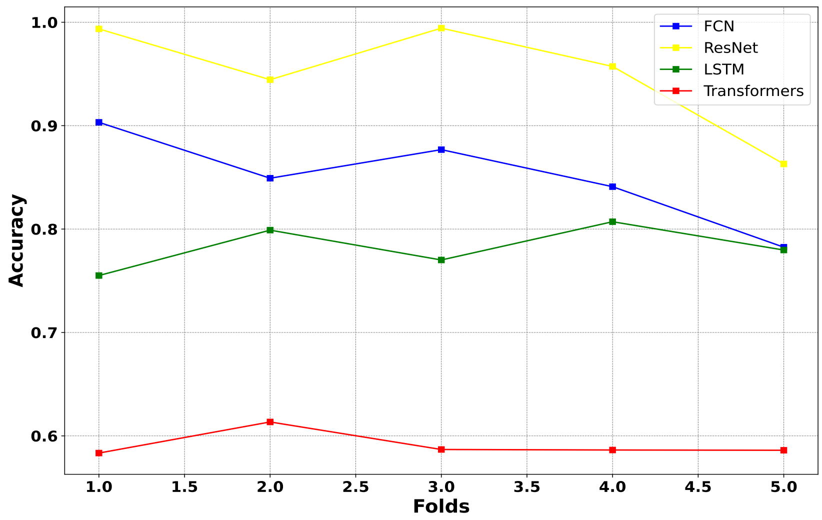


Figure 5.1: *KFold (k=5) prediction accuracy on different models*

of the data readings obtained from Tobii 4C-Pro eye trackers lack data. Consequently, the eye tracker data was not used for training our models. To ensure accurate results, we decided to focus solely on the physiological data and avoid using the eye tracker data. This allowed us to obtain reliable data for our models. The results of every model implemented are shown in the next subsection.

5.1 Classification with FCN, ResNet, Transformers and LSTM

Lacking fine-tuning, the FCN model achieved an accuracy rate of 41.16% and an F1 score of 0.483. However, notable improvements were observed after incorporating the fine-tuning process. 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. A considerable shift in

Table 5.1: *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**

average accuracy from no fine-tuning to 20% fine-tuning of test data can was noticed. The results showed that the FCN model was able to effectively learn from fine-tuned data. The model accuracy improved by nearly 45% after fine-tuning the model with 20% test data. This highlights the importance of fine-tuning in the development of deep learning models. Additionally, using the KFold approach, the FCN model was able to classify the stress levels with an average accuracy rate of 85.05% and an average F1 score of 0.848. Each fold in KFold cross-validation of FCN maintains an accuracy between

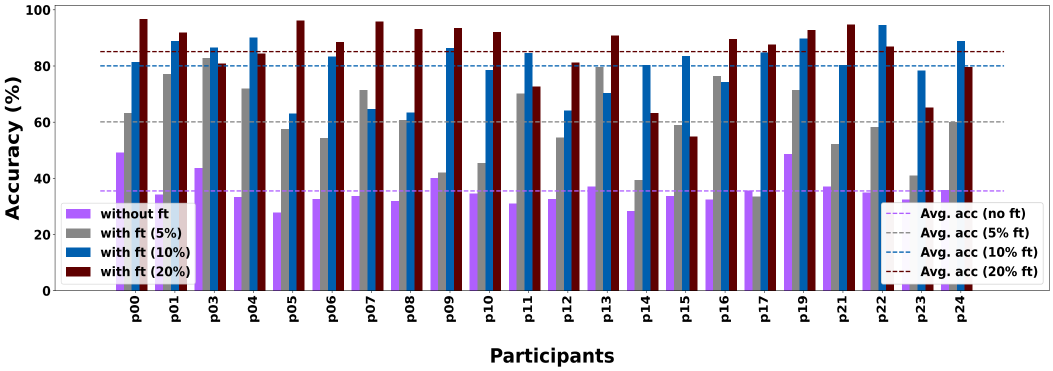


Figure 5.2: *ResNet accuracy on LOPO with and without fine-tuning*

75% and 90% as shown in Figure 5.1.

Figure 5.2 shows ResNet accuracy on LOPO with and without fine-tuning. With no fine-tuning technique, this densely constructed model classified the three classes with an accuracy of 35.52%. However, a significant increase in accuracy rate to 60.1% (F1 score = 0.555) was seen after fine-tuning the model with only 5% of the testing data from each participant. Likewise, ResNet made a significant improvement in predicting the labels when further fine-tuned with 10% and 20% test data with accuracy rates of 80.04% and 85.12% and F1 scores of 0.786 and 0.846, respectively.

Similarly, the sampled EDA, BVP, TEMP and ACC data were utilized in the implementation of Transformers. Transformers achieved an accuracy rate of 40.38% and an F1 score of 0.375 with LOPO, with no fine-tuning. Despite the initial accuracy achieved by Transformers, the fine-tuning process did not yield a notable improvement in its classification capabilities. The model did not experience a substantial change in performance with the application of the fine-tuning method. However, the individual participants showed a significant improvement in accuracy when treating the training data with 10% of the testing data. This suggests that the fine-tuning process had a

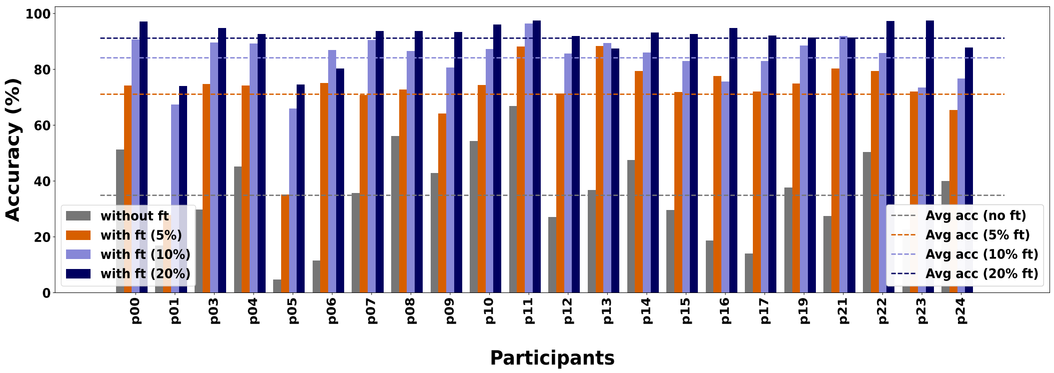


Figure 5.3: *LSTM accuracy on LOPO with and without fine-tuning*

positive effect on the model's ability to generalize. By maintaining prediction accuracy between 58% and 61%, Transformers achieved an average accuracy rate of 59.12% and an average F1 score of 0.57. Similar to LOPO with and without fine-tuning, KFold on Transformers did not perform up to the mark compared to other deep learning models. However, Transformers still achieved slightly better accuracy than LOPO. The results suggest that Transformers can still be used to achieve reasonable performance on classification tasks.

LSTM has achieved remarkable results compared to other models implemented. Same as Transformers, the sampled physiological signals were taken to train and validate the model. Without fine-tuning, the LSTM model achieved an accuracy of 35.86% in classifying 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. By examining the bar chart in Figure 5.3, one can assess the impact of fine-tuning on the model's ability to accurately classify and distribute stress instances in a generalized as well as a user-centric way. KFold on LSTM performed better than LOPO with no fine-tuning and LOPO with 5% fine-tuning in stress level classification. It classified the three classes with an average accuracy of 78.21% and an F1 score of 0.781. However, it was observed that KFold had a slightly lower impact in comparison to the fine-tuning techniques of 10% and 20%. This indicates that KFold was still able to capture the subtle changes in the data better than the fine-tuning techniques, although the fine-tuning techniques had a slightly higher impact due to their ability to 'tune' the model to the participant-specific data.

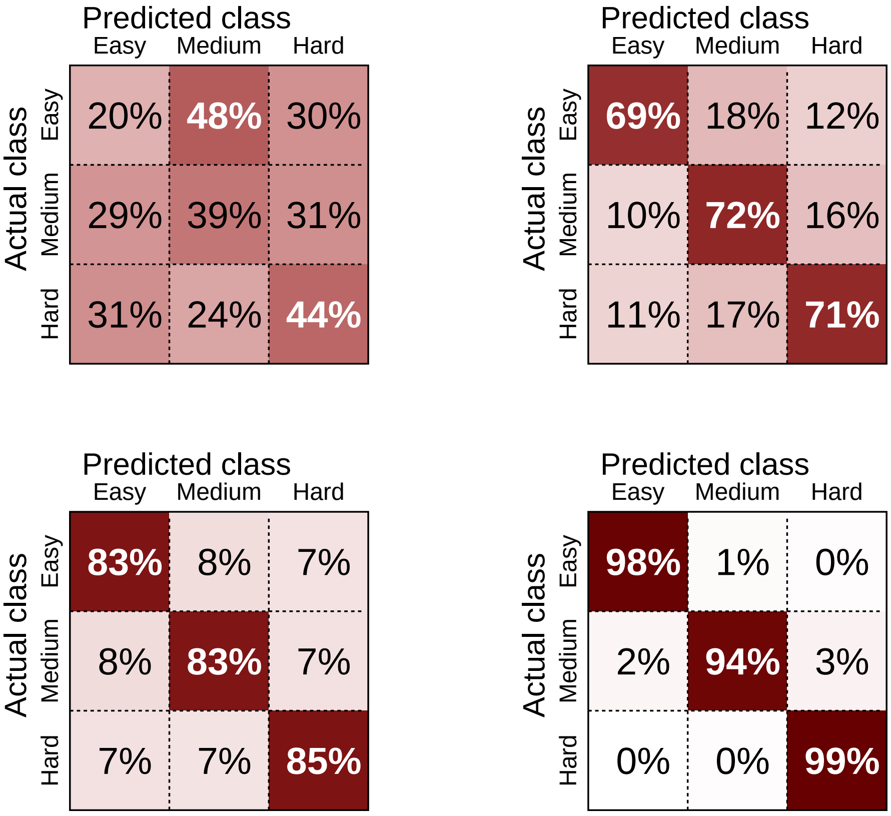


Figure 5.4: *Confusion matrix of LSTM without fine-tuning (top-left), with 5% fine-tuning (top-right), 10% fine-tuning (bottom-left) and 20% fine-tuning (bottom-right).*

5.2 Comparison with confusion matrix

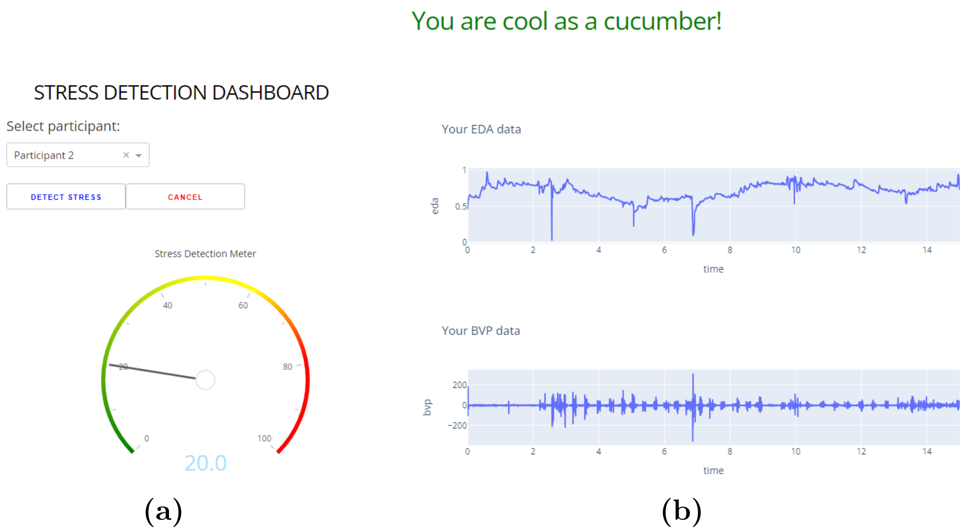
The confusion matrices for LSTM are displayed in Figure 5.4, which illustrates the changes in classification performance when fine-tuning is applied compared to when it is not. These matrices provide a visual representation of how well the model performs in accurately predicting the three distinct stress classes: easy, medium and hard. By examining the confusion matrices, we can observe the distribution of predicted and actual stress classes and identify how the fine-tuning process enhanced classification performance even with a small percentage of data. Overall, the models were able to accurately classify the three stress classes with a high degree of accuracy using fine-tuning.

Chapter 6

User Application

In addition to implementing the models, we developed a prototype application using Python Dash to visualize and represent the variations in stress levels observed by each participant in a timely manner. The application interface, as illustrated in Figure 6.1, was designed to provide a comprehensive overview of stress levels over time. This was done using the FCN model implemented with user-specific data. We used a training method tailored to each individual, where we split the data of each participant into training and testing sets and use the test set to validate the stress levels. This allowed the stress visualization application to accurately capture and display each person's unique characteristics and differences. We then utilized the trained model to predict each user's stress level and visualize it in the application.

Additionally, this approach was found to be more effective than using a generic model for all users. The results showed improved accuracy and performance in the visualization of stress levels. By implementing this personalized training methodology, we aim to improve the overall personalization and distinctiveness of the application for each user, as a standard model may not account for individual differences. This will enable



You are cool as a cucumber!

STRESS DETECTION DASHBOARD

Select participant: Participant 2

(a) (b)

Figure 6.1: *Screenshot of feedback application (a) Stress-meter (b) Customized message with EDA and BVP signals*

the application to provide more targeted and effective training solutions for users and improve the user experience. Additionally, this will also help to ensure the accuracy and precision of the application's data analysis.

Within the application, users had the flexibility to select participants from a list for analysis. Once a participant was chosen, the application showed the dynamic changes in the stress meter. It also displayed 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 allowed users 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, leading to 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.

By providing real-time insights into student's mental state while performing various tasks, students will be able to understand, manage and address their stress levels effectively. A tool like this could be extremely effective in aiding student's well-being and enhancing their academic achievement by promoting self-awareness and the ability to manage stress in a proactive manner. This would also help them develop the necessary skills to effectively cope with stress in order to achieve their desired goals. Additionally, this kind of tool could be used as a complement to traditional educational techniques, helping students to better understand their own emotional states. By doing so, it could help students become more productive in their daily activities, have better relationships with their peers and be better prepared to face their future. The tool could be used as a way to get feedback from students and provide personalized coaching to help them make better decisions and reach their goals. Ultimately, this could lead to a more positive learning experience and better outcomes.

Chapter 7

Discussion

In this section, the noteworthy contributions mentioned in Chapter 1 are discussed along with the findings and responses to our research questions. Additionally, some of the limitations and future work of our research are also discussed in the following subsections.

7.1 Achieving Distinctiveness in Data Collection

As part of our research, we explored how a student's physiological responses vary under different levels of stress. Our objective was to gain a better understanding of how stress impacts the well-being and performance of students. We discovered that students experienced an increase in physiological reactions as their stress levels intensified. Our aim was to demonstrate that stress can have a detrimental impact on physiological responses, leading to negative effects on a student's performance and well-being. To confirm these findings, we conducted an experiment to measure stress levels in students. We used mental arithmetic tasks with feedback, buzzers and timers with clock-ticking sounds to create extra stress and obtain distinct variations. We designed a controlled system with preset ground truths to eliminate potential biases and ensure an objective assessment of the data. By using a multilevel approach, our classification system provides a detailed and comprehensive understanding of different levels of stress experienced by individuals. This enables us to develop more personalized and effective stress management and intervention approaches. Our ultimate goal is to create better strategies for managing and mitigating stress and this model is a crucial step towards achieving that goal. By understanding the different levels of stress, we are able to identify which individuals may be more vulnerable to the adverse effects of stress. This helps us to create more targeted interventions to better manage and reduce stress levels. Additionally, by being able to identify the different levels of stress, we can better understand how stress impacts the individual and the community as a whole, allowing us to develop more effective strategies for tackling the issue.

We collected data from 25 participants across four stress levels using wristbands and eye trackers. Participants were split into two groups, with one group starting with relaxation and the other ending with it. To determine if the timing of the 'relax' session had any effect on stress levels among the two groups, this approach tested the potential effect of the 'relax' session on participant stress levels. However, our analysis did not show any significant differences between the two groups. Additionally, we encountered some difficulties during the experiment as some participants had difficulty concluding the E4 recording after the 'relax' session. This led to inaccuracies in labeling data and resulted in its exclusion from model training. On the other hand, the data from the eye tracker were also discarded from being incorporated into the analysis. To consider the eye data, the experiment would have included a design in such a way that participants do not take their eye away from the screen or simply use a head-mounted eye tracker.

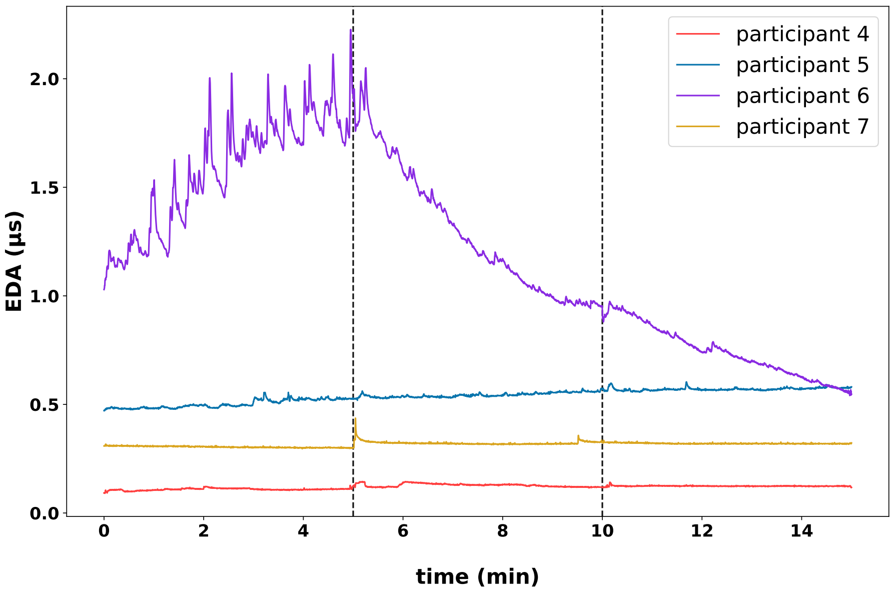


Figure 7.1: *Physiological responses from the four participants. (x-axis: 0 to 5 - hard level, 5 to 10 - medium level, 10 to 15 - easy level)*

The new design would allow for a more sophisticated method of entering answers without using a keyboard. Instead, it would be easier to select the answers on the screen with multiple options. Eye data would provide more accurate results and allow for a better understanding of user behavior. Additionally, the data collected would be more reliable and lead to more precise insights.

7.2 Deep learning models to predict stress levels

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 key consideration was heart rate (HR) data, which can be derived from blood volume pulse (BVP) data sampled at 64Hz [81]. Given this relationship, HR data was excluded 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 were down-sampled to 4Hz. Unfortunately, downsampling resulted in data loss from the BVP and ACC sensors. On the other hand, the FCN and ResNet implementations utilized all the raw data at 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. This showcases the efficacy of the chosen approaches in leveraging deep learning models to tackle the complex task of stress level classification using E4 wristband data.

7.3 User-independent vs User-dependent prediction analysis

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 as shown in Figure 7.1. To address these inter-participant variations, the models were calibrated using a small percentage of the test data through a fine-tuning process. Interestingly, this fine-tuning procedure significantly improved classification performance. By incorporating just 5% of person-specific data for training purposes, a

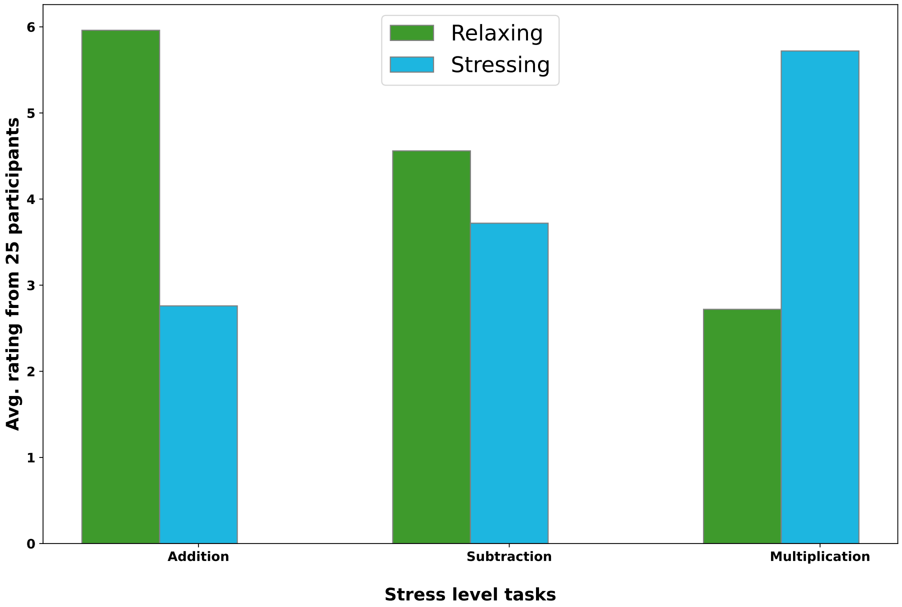


Figure 7.2: *Post-questionnaire responses from all participants indicating the easy and difficult levels*

significant improvement in performance was observed. This improvement underscores the importance of fine-tuning or calibrating 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 individual differences. This means that the use of personalized data in the calibration of stress models can improve the accuracy and reliability of stress classification. It is therefore essential to consider individual differences when designing stress classification algorithms.

7.4 Correlation between the ground truths and feedback from participants

Our experiment included the collection of feedback from participants before and after the main sessions. Even though this data is not used for our evaluation of physiological signals, it is interesting to note that there is a correlation between the ground truths and feedback from participants. Figure 7.2 shows the feedback from participants for our post-questionnaire form, where participants were asked about the easy, medium and difficult tasks in the experiment. Based on our ground truths, addition was easy, subtraction was moderate and multiplication was more challenging. The same can be found in the figure. The bar chart shows that the participants had a similar understanding of the stress level during the experiment. Furthermore, this also demonstrates the effectiveness of our experimental design and its ability to differentiate the difficulty of the tasks.

On the other hand, it is misleading to rely completely on the responses or feedback given by the participants on the questionnaire forms to evaluate the model system. This is because the responses might be biased or influenced by the participant's subjective opinions. Moreover, the responses might not reflect the actual performance of the participants. For example, the post-questionnaire feedback from participants *p04, p05, p06* and *p07*, who were in the same group performing tasks in order, hard->medium->easy, selected addition as the least difficult task and multiplication as the most difficult task. Figure 7.1 illustrates the physiological responses (EDA) from the same four participants, where only participant *p06* exhibited physiological responses as described in the feedback. In the first 5 min (hard level), the EDA signals are rising and at their peak, while in the next 10 min (medium and easy level), the EDA responses are gradually decreasing indicating a decrease in stress. However, the psychological responses of the participants, *p04, p05* and *p07*, differ slightly from their feedback on the questionnaires. The EDA signals from these three participants do not show a heavy rise in hard level and a gradual decrease in medium and easy levels. In fact, the participant p05 exhibits a contrasting behavior where the signals keep increasing from hard to easy level. It is therefore critical to combine qualitative and quantitative data to assess the system's performance accurately.

Chapter 8

Conclusion and Future work

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 the 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 of LOPO. Overall, the study findings demonstrate the effectiveness of the deep learning-based approach to predicting stress levels using physiological signals. As a result of the predictions, we have found that ResNet and LSTM models are superior in terms of accuracy and fine-tuning or calibration techniques are beneficial in enhancing the accuracy of generalized predictions related to stress levels. This research underscores the power of deep learning in stress analysis and offers valuable insights for future studies in this field. Additionally, a prototype application was developed to visually display stress levels and provide customized alerts for timely support.

Expanding the scope of this study to encompass a real-time stress detection system could have transformative implications in education and healthcare. Such a system has the potential to make 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-induced 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. It is imperative that we take advantage of this technology to advance our knowledge and abilities in this crucial area. This technology has the potential to help individuals and organizations better understand their stress levels and develop better strategies for managing stress. It can also provide valuable insights into how to better design work environments that promote healthy and productive work. Finally, it can aid in the research and development of new treatments and interventions for stress-related issues.

Bibliography

[1] Katie-Jane Brickwood, Greig Watson, Jane O'Brien, Andrew D Williams, et al. Consumer-based wearable activity trackers increase physical activity participation: systematic review and meta-analysis. *JMIR mHealth and uHealth*, 7(4):e11819, 2019.

[2] Frank L Schwartz, Cynthia R Marling, and Razvan C Bunescu. The promise and perils of wearable physiological sensors for diabetes management. *Journal of diabetes science and technology*, 12(3):587–591, 2018.

[3] Ankita Agarwal, Josephine Graft, Noah Schroeder, and William Romine. Sensor-based prediction of mental effort during learning from physiological data: A longitudinal case study. *Signals*, 2(4):886–901, 2021.

[4] ZX Zhou, Vincent Tam, King-Shan Lui, Edmund Y Lam, Allan Yuen, Xiao Hu, and Nancy Law. Applying deep learning and wearable devices for educational data analytics. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 871–878. IEEE, 2019.

[5] Yu Shi, Natalie Ruiz, Ronnie Taib, Eric Choi, and Fang Chen. Galvanic skin response (gsr) as an index of cognitive load. In *CHI'07 extended*

abstracts on Human factors in computing systems, pages 2651–2656, 2007.

[6] Hindra Kurniawan, Alexandr V Maslov, and Mykola Pechenizkiy. Stress detection from speech and galvanic skin response signals. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, pages 209–214. IEEE, 2013.

[7] Siao Zheng Bong, M Murugappan, and Sazali Yaacob. 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*, pages 198–205. Springer, 2012.

[8] Martin Gjoreski, Hristijan Gjoreski, Mitja Luštrek, and Matjaž Gams. 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*, pages 1185–1193, 2016.

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

[10] Kazunori Ushiyama, Takeshi Ogawa, Masanori Ishii, Ryuichi Ajisaka, Yasuro Sugishita, and Iwao Ito. Physiologic neuroendocrine arousal by mental arithmetic stress test in healthy subjects. *The American journal of cardiology*, 67(1):101–103, 1991.

[11] Fred Paas and John Sweller. Implications of cognitive load theory for

multimedia learning. *The Cambridge handbook of multimedia learning*, 27:27–42, 2014.

[12] Aamir Arsalan, Syed Muhammad Anwar, and Muhammad Majid. Mental stress detection using data from wearable and non-wearable sensors: a review. *arXiv preprint arXiv:2202.03033*, 2022.

[13] Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. A global measure of perceived stress. *Journal of health and social behavior*, pages 385–396, 1983.

[14] Phillip J Brantley, Craig D Waggoner, Glenn N Jones, and Neil B Rappaport. A daily stress inventory: Development, reliability, and validity. *Journal of behavioral medicine*, 10:61–73, 1987.

[15] Leonard R Derogatis. *Brief symptom inventory: BSI.* Pearson, 1993.

[16] Richard A Bryant, Michelle L Moulds, and Rachel M Guthrie. Acute stress disorder scale: a self-report measure of acute stress disorder. *Psychological assessment*, 12(1):61, 2000.

[17] Ingun Ulstein, Torgeir Bruun Wyller, and Knut Engedal. The relative stress scale, a useful instrument to identify various aspects of carer burden in dementia? *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*, 22(1):61–67, 2007.

[18] Roland Brunken, Jan L Plass, and Detlev Leutner. Direct measurement of cognitive load in multimedia learning. *Educational psychologist*, 38(1):53–61, 2003.

[19] Guanghua Wu, Guangyuan Liu, and Min Hao. The analysis of emotion recognition from gsr based on pso. In *2010 International symposium on*

intelligence information processing and trusted computing, pages 360–363. IEEE, 2010.

[20] Fitbit. https://www.fitbit.com/global/us/products/.

[21] Emotiv Epoc X - 14 channel wireless EEG Headset. https://www.emotiv.com/epoc-x/.

[22] Ahmad Rauf Subhani, Wajid Mumtaz, Mohamed Naufal Bin Mohamed Saad, Nidal Kamel, and Aamir Saeed Malik. Machine learning framework for the detection of mental stress at multiple levels. *IEEE Access*, 5:13545–13556, 2017.

[23] MM Sani, H Norhazman, HA Omar, Norliza Zaini, and SA Ghani. Support vector machine for classification of stress subjects using eeg signals. In *2014 IEEE Conference on Systems, Process and Control (ICSPC 2014)*, pages 127–131. IEEE, 2014.

[24] Joong Woo Ahn, Yunseo Ku, and Hee Chan Kim. A novel wearable eeg and ecg recording system for stress assessment. *Sensors*, 19(9):1991, 2019.

[25] Joachim Taelman, Steven Vandeput, Ivan Gligorijević, Arthur Spaepen, and Sabine Van Huffel. Time-frequency heart rate variability characteristics of young adults during physical, mental and combined stress in laboratory environment. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1973–1976. IEEE, 2011.

[26] P Karthikeyan, M Murugappan, and Sazali Yaacob. Ecg signals based mental stress assessment using wavelet transform. In *2011 IEEE International Conference on Control System, Computing and Engineering*, pages 258–262. IEEE, 2011.

[27] Mind Media. https://www.mindmedia.com/en/solutions/research/bloodvolume-pulse-ppg/.

[28] P Karthikeyan, M Murugappan, and Sazali Yaacob. 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*, pages 236–243. Springer, 2012.

[29] Enrique Garcia-Ceja, 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):1053–1060, 2015.

[30] Empatica. E4 wristband. https://www.empatica.com/en-int/research/e4/.

[31] Tobii. https://www.digitaltrends.com/gaming/tobii-eye-tracker-4c-announced/.

[32] Javier Hernandez, Rob R Morris, and Rosalind W Picard. Call center stress recognition with person-specific models. In *Affective Computing and Intelligent Interaction: 4th International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part I 4*, pages 125–134. Springer, 2011.

[33] Saroj Kumar Panigrahy, Sanjay Kumar Jena, and Ashok Kumar Turuk. Study and analysis of human stress detection using galvanic skin response (gsr) sensor in wired and wireless environments. *Research Journal of Pharmacy and Technology*, 10(2):545–550, 2017.

[34] Tong Boon Tang, Lip Wee Yeo, and Dandy Jing Hui Lau. Activity

awareness can improve continuous stress detection in galvanic skin response. In *SENSORS, 2014 IEEE*, pages 1980–1983. IEEE, 2014.

[35] Xiyuan Hou, Yisi Liu, Olga Sourina, and Wolfgang Mueller-Wittig. Cognimeter: Eeg-based emotion, mental workload and stress visual monitoring. In *2015 International Conference on Cyberworlds (CW)*, pages 153–160. IEEE, 2015.

[36] Rossana Castaldo, William Xu, Paolo Melillo, Leandro Pecchia, Lorena Santamaria, and C James. Detection of mental stress due to oral academic examination via ultra-short-term hrv analysis. In *2016 38th Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 3805–3808. IEEE, 2016.

[37] Somchanok Tivatansakul and Michiko Ohkura. Improvement of emotional healthcare system with stress detection from ecg signal. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6792–6795. IEEE, 2015.

[38] Giorgia Acerbi, Erika Rovini, Stefano Betti, Antonio Tirri, Judit Flóra Rónai, Antonella Sirianni, Jacopo Agrimi, Lorenzo Eusebi, and Filippo Cavallo. A wearable system for stress detection through physiological data analysis. In *Ambient Assisted Living: Italian Forum 2016 7*, pages 31–50. Springer, 2017.

[39] C Schubert, M Lambertz, RA Nelesen, W Bardwell, J-B Choi, and JE Dimsdale. Effects of stress on heart rate complexity—a comparison between short-term and chronic stress. *Biological psychology*, 80(3):325–332, 2009.

[40] Daniel J McDuff, Javier Hernandez, Sarah Gontarek, and Rosalind W Picard. Cogcam: Contact-free measurement of cognitive stress during

computer tasks with a digital camera. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4000–4004, 2016.

[41] Muhammad Zubair and Changwoo Yoon. Multilevel mental stress detection using ultra-short pulse rate variability series. *Biomedical Signal Processing and Control*, 57:101736, 2020.

[42] Andrew Steptoe, Gonneke Willemsen, Natalie Owen, Louise Flower, and Vidya Mohamed-Ali. Acute mental stress elicits delayed increases in circulating inflammatory cytokine levels. *Clinical Science*, 101(2):185–192, 2001.

[43] Nis Hjortskov, Dag Rissén, Anne Katrine Blangsted, Nils Fallentin, Ulf Lundberg, and Karen Søgaard. The effect of mental stress on heart rate variability and blood pressure during computer work. *European journal of applied physiology*, 92:84–89, 2004.

[44] Christiaan H Vinkers, Renske Penning, Juliane Hellhammer, Joris C Verster, John HGM Klaessens, Berend Olivier, and Cor J Kalkman. The effect of stress on core and peripheral body temperature in humans. *Stress*, 16(5):520–530, 2013.

[45] Palanisamy Karthikeyan, Murugappan Murugappan, and Sazali Yaacob. Descriptive analysis of skin temperature variability of sympathetic nervous system activity in stress. *Journal of Physical Therapy Science*, 24(12):1341–1344, 2012.

[46] Mikhail Sysoev, Andrej Kos, and Matevž Pogačnik. Noninvasive stress recognition considering the current activity. *Personal and Ubiquitous Computing*, 19:1045–1052, 2015.

[47] Martin Gjoreski, Hristijan Gjoreski, Mitja Lutrek, and Matja Gams. Automatic detection of perceived stress in campus students using smartphones. In *2015 International conference on intelligent environments*, pages 132–135. IEEE, 2015.

[48] Martijn Haak, Steven Bos, Sacha Panic, and Léon JM Rothkrantz. Detecting stress using eye blinks and brain activity from eeg signals. *Proceeding of the 1st driver car interaction and interface (DCII 2008)*, pages 35–60, 2009.

[49] HM Simpson and FM Molloy. Effects of audience anxiety on pupil size. *Psychophysiology*, 8(4):491–496, 1971.

[50] Luma Tabbaa, Ryan Searle, Saber Mirzaee Bafti, Md Moinul Hossain, Jittrapol Intarasisrisawat, Maxine Glancy, and Chee Siang Ang. Vreed: Virtual reality emotion recognition dataset using eye tracking & physiological measures. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 5(4):1–20, 2021.

[51] LA Torres-Salomao, Mahdi Mahfouf, and Emad El-Samahy. Pupil diameter size marker for incremental mental stress detection. In *2015 17th international conference on e-health networking, application & services (HealthCom)*, pages 286–291. IEEE, 2015.

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

[53] Joe Tomaka, Jim Blascovich, and Laura Swart. Effects of vocalization on cardiovascular and electrodermal responses during mental arithmetic. *International Journal of Psychophysiology*, 18(1):23–33, 1994.

[54] Katarina Dedovic, Robert Renwick, Najmeh Khalili Mahani, Veronika Engert, Sonia J Lupien, and Jens C Pruessner. The montreal imaging stress task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain. *Journal of Psychiatry and Neuroscience*, 30(5):319–325, 2005.

[55] Cornelia Setz, Bert Arnrich, Johannes Schumm, Roberto La Marca, Gerhard Tröster, and Ulrike Ehlert. Discriminating stress from cognitive load using a wearable eda device. *IEEE Transactions on information technology in biomedicine*, 14(2):410–417, 2009.

[56] Bert Arnrich, Cornelia Setz, Roberto La Marca, Gerhard Tröster, and Ulrike Ehlert. What does your chair know about your stress level? *IEEE Transactions on Information Technology in Biomedicine*, 14(2):207–214, 2009.

[57] J Ridley Stroop. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6):643, 1935.

[58] Stefan Koelsch and Thomas Stegemann. The brain and positive biological effects in healthy and clinical populations. *Music, health, and wellbeing*, pages 436–456, 2012.

[59] Stephanie Khalfa, SIMONE DALLA BELLA, Mathieu Roy, Isabelle Peretz, and Sonia J Lupien. Effects of relaxing music on salivary cortisol level after psychological stress. *Annals of the New York Academy of Sciences*, 999(1):374–376, 2003.

[60] Anum Asif, Muhammad Majid, and Syed Muhammad Anwar. Human stress classification using eeg signals in response to music tracks. *Computers in biology and medicine*, 107:182–196, 2019.

[61] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H Hellhammer. The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2):76–81, 1993.

[62] Mariette Soury and Laurence Devillers. Stress detection from audio on multiple window analysis size in a public speaking task. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 529–533. IEEE, 2013.

[63] Virginia Sandulescu, Sally Andrews, David Ellis, Nicola Bellotto, and Oscar Martinez Mozos. Stress detection using wearable physiological sensors. In *Artificial Computation in Biology and Medicine: International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2015, Elche, Spain, June 1-5, 2015, Proceedings, Part I 6*, pages 526–532. Springer, 2015.

[64] Bruce S McEwen. Glucocorticoids, depression, and mood disorders: structural remodeling in the brain. *Metabolism*, 54(5):20–23, 2005.

[65] Shelley E Taylor, Laura Cousino Klein, Brian P Lewis, Tara L Gruenewald, Regan AR Gurung, and John A Updegraff. Biobehavioral responses to stress in females: tend-and-befriend, not fight-or-flight. *Psychological review*, 107(3):411, 2000.

[66] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. International affective picture system (iaps): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 1(39-58):3, 1997.

[67] Seyyed Abed Hosseini, Mohammad Ali Khalilzadeh, Mohammad Bagher Naghibi-Sistani, and Vahid Niazmand. Higher order spectra analysis

of eeg signals in emotional stress states. In *2010 Second international conference on information technology and computer science*, pages 60–63. IEEE, 2010.

[68] George Tanev, Dorthe B Saadi, Karsten Hoppe, and Helge BD Sorensen. Classification of acute stress using linear and non-linear heart rate variability analysis derived from sternal ecg. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3386–3389. IEEE, 2014.

[69] David A Grant and Esta A Berg. Wisconsin card sorting test. *Journal of Experimental Psychology*, 1993.

[70] Boon-Giin Lee and Wan-Young Chung. Wearable glove-type driver stress detection using a motion sensor. *IEEE Transactions on Intelligent Transportation Systems*, 18(7):1835–1844, 2016.

[71] Saskia Koldijk, Maya Sappelli, Suzan Verberne, Mark A Neerincx, and Wessel Kraaij. The swell knowledge work dataset for stress and user modeling research. In *Proceedings of the 16th international conference on multimodal interaction*, pages 291–298, 2014.

[72] Valentina Markova, Todor Ganchev, and Kalin Kalinkov. Clas: A database for cognitive load, affect and stress recognition. In *2019 International Conference on Biomedical Innovations and Applications (BIA)*, pages 1–4. IEEE, 2019.

[73] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[74] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline.

In *2017 International joint conference on neural networks (IJCNN)*, pages 1578–1585. IEEE, 2017.

[75] Maciej Dzieżyc, Martin Gjoreski, Przemysław Kazienko, Stanisław Saganowski, and Matjaž Gams. Can we ditch feature engineering? end-to-end deep learning for affect recognition from physiological sensor data. *Sensors*, 20(22):6535, 2020.

[76] Behnam Behinaein, Anubhav Bhatti, Dirk Rodenburg, Paul Hungler, and Ali Etemad. A transformer architecture for stress detection from ecg. In *Proceedings of the 2021 ACM International Symposium on Wearable Computers*, pages 132–134, 2021.

[77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[78] Theodoros Ntakouris. `https://keras.io/examples/timeseries/`, 2021.

[79] Jürgen Schmidhuber, Sepp Hochreiter, et al. Long short-term memory. *Neural Comput*, 9(8):1735–1780, 1997.

[80] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[81] Erik Peper, Rick Harvey, I-Mei Lin, Hana Tylova, and Donald Moss. Is there more to blood volume pulse than heart rate variability, respiratory sinus arrhythmia, and cardiorespiratory synchrony? *Biofeedback*, 35(2), 2007.