Stress Detection Through Wrist-Based Electrodermal Activity Monitoring and Machine Learning

Lili Zhu, *Graduate Student Member, IEEE*, Petros Spachos, *Senior Member, IEEE*, Pai Chet Ng, Yuanhao Yu, Yang Wang, Konstantinos Plataniotis, *Fellow, IEEE*, and Dimitrios Hatzinakos, *Fellow, IEEE*

***Abstract*—Stress is an inevitable part of modern life. While stress can negatively impact a person's life and health, positive and under-controlled stress can also enable people to generate creative solutions to problems encountered in their daily lives. Although it is hard to eliminate stress, we can learn to monitor and control its physical and psychological effects. It is essential to provide feasible and immediate solutions for more mental health counselling and support programs to help people relieve stress and improve their mental health. Popular wearable devices, such as smartwatches with several sensing capabilities, including physiological signal monitoring, can alleviate the problem. This work investigates the feasibility of using wrist-based electrodermal activity (EDA) signals collected from wearable devices to predict people's stress status and identify possible factors impacting stress classification accuracy. We use data collected from wrist-worn devices to examine the binary classification discriminating stress from non-stress. For efficient classification, five machine learning-based classifiers were examined. We explore the classification performance on four available EDA databases under different feature selections. According to the results, Support Vector Machine (SVM) outperforms the other machine learning approaches with an accuracy of 92.9 for stress prediction. Additionally, when the subject classification included gender information, the performance analysis showed significant differences between males and females. We further examine a multimodal approach for stress classifications. The results indicate that wearable devices with EDA sensors have a great potential to provide helpful insight for improved mental health monitoring.**

***Index Terms*—Stress measurement, emotion recognition, EDA, wearable sensors, wrist-worn wearable device, smartwatches, k-nearest neighbors, support vector machines, naive bayes, logistic regression, random forest.**

I. Introduction

With the rapid development of modern society, people's understanding of health is no longer limited to physical health but extends to open concepts such as mental health and social adaptation. With good mental health, people can have a solid attitude to fulfill family and social responsibilities, actively find solutions to problems, and positively plan for the future.

Currently, medical workers and researchers focus on interventions and treatments for existing mental issues and symptoms. However, less-obvious stress and anxiety should also be noticed. Before stress is high enough to alert people to visit a doctor, the body and brain are already sending signals to help people know what needs to be changed. These signals include the inability to concentrate, memory loss, unstable emotions, impatient, restlessness, etc. When people experience these physical and mental reactions to stress, their physiological signals also change. For instance, in [1], they found that stress responses such as respiration rate, heart rate, and electrodermal activity (EDA) signals increased when a sound of an air horn rang suddenly. Therefore, learning the corresponding relationships between the changes in different physiological signals and stress is significant for identifying stress. Meanwhile, utilizing objective scientific and technological means to help monitor the physiological signals and analyze the responses has become an important research topic in academia and industry.

Wearable technology is a promising solution for remote and continuous mental health monitoring. Wearable devices can capture rich contextual information and deliver a large amount of personal patient data. At the same time, the advantages of machine learning have increased data processing speed and provided better data insights. Among the popular wearable devices are smartwatches, that act as mini smartphones with promising computational capabilities, while they also have many sensors that can collect physiological signals, including EDA, Photoplethysmography (PPG), Electrocardiogram (ECG), and skin temperature.

This work examines the feasibility of using EDA data collected from wrist-worn wearable devices to detect human stress.

Manuscript received 25 February 2022; revised 7 July 2022 and 23 December 2022; accepted 11 January 2023. Date of publication 23 January 2023; date of current version 5 May 2023. This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada under Grant RGPIN/2016-04007. *(Corresponding author: Petros Spachos.)*

Lili Zhu and Petros Spachos are with the School of Engineering, University of Guelph, Guelph, ON N1G 2W1, Canada (e-mail: lzhu03@uoguelph.ca; petros@uoguelph.ca).

Pai Chet Ng, Konstantinos Plataniotis, and Dimitrios Hatzinakos are with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: pc.ng@utoronto.ca; kostas@ece.utoronto.ca; dimitris@ece.utoronto.ca).

Yuanhao Yu and Yang Wang are with the Huawei Technologies Canada, ON L3R 5A4, Canada (e-mail: yuanhao.yu@huawei.com; yang.wang3@huawei.com).

Digital Object Identifier 10.1109/JBHI.2023.3239305

EDA can be captured through a sensor that measures skin conductance changes. Since skin conductance can reflect the human body's emotions and physiological responses, EDA is often used as a physiological indicator to measure emotional changes. Some commercially available smartwatches and wearable devices (e.g., Fitbit sense, Empatica E4) already have integrated EDA sensors to deliver emerging applications, such as emotion monitoring to prevent excessive tension and anxiety [2]. However, in contrast to the medical-grade EDA sensors that need to be at a specific position stationary, wearable devices, especially smartwatches, are always worn on a human's wrist with uncertain motion. Such dynamic movement by humans creates an elusive challenge to detect human stress with EDA sensors from wearable devices.

To verify the feasibility of EDA data for the above purposes, we explore four public datasets that provide EDA signals collected from wearable devices. Next, we applied five machine learning methods to compare their performances on classifying stress and non-stress status. We present the classification results of training with all features and extracted features. We further examine any differences between males and females regarding the relationship between stress and EDA, and possible reasons are analyzed. Finally, we discuss a multimodal approach for wrist-worn wearable devices and stress detection.

The main contributions of this paper are summarized below:

- Stress detection is feasible through EDA signals collected from wrist-worn wearable devices. In all the datasets, EDA provides an accuracy above 70% for stress classification.
- The proposed system needs to be executable on smartphones and smartwatches so that simple machine learning methods are examined. Among the discussed methods, SVM achieved the highest accuracy of 92.9% in one of the datasets.
- Compared to other modalities, including PPG and ECG, and different combinations, EDA provides the highest accuracy in stress classification.
- When gender information is available, EDA stress classification accuracy is higher in females, and for the datasets that we used.

The rest of this paper is organized as follows: Section II illustrates a literature review on the works related to this study. Section III introduces the system overview of this work and related concepts. Section IV presents the methods we adopted for the experiment. Section V analyzes the performances of the methods and the observations from the results. Finally, Section VI summarizes this work.

II. Related Works

EDA signals can be used as an indicator of emotional change. EDA data are used to monitor, analyze and evaluate people's emotional and stress responses in many scenarios. In [3], they experimented on whether EDA is a helpful indicator of emotional reactions when working individually, cooperating, and competing with others. The experimental results revealed that, in collaborative tasks, participants produced more distinguished EDA signals than in competitive tasks. In addition, when males and females were engaged in different tasks, skin conductance level (SCL) and non-skin conductance responses (NS-SCRs) showed different patterns. In [4], they discovered a significant relationship between EDA signals and the self-reported arousal scores of participants who read emotional content loudly. In [5], they confirmed that EDA could reflect social discomfort by asking twenty-eight participants to take a radial line bisection task individually or with a stranger. The outcomes showed that the EDA fluctuations and performances of the participants who had physical discomfort and those who did not were distinctive. The feasibility of adopting EDA to identify people's stress arousal when they are underwater has been studied in [6].

Many works exploit the use of EDA in different scenarios targeting different groups of users. For example, EDA was used to estimate how workers perceived potential risks when performing construction work [7]. In [8], EDA was applied to recognize patients' anxiety about surgery. In [9], they used ECG and EDA data to check the driver's stress reactions when driving in a simulator with several car settings. In [10], again with drivers, they used Fisher projection and linear discriminant analysis to detect drivers' stress levels based on EDA under different driving conditions, and the methods had a recognition rate of 81.82%. In [11], they summarized that adolescents with major depressive disorder (MDD) had significantly low EDA while continuous recording, indicating that the sympathetic part of the autonomic nervous system of adolescents with MDD is dysregulated.

Since the global spreading of the COVID-19 pandemic, several aspects of people's lives have undergone tremendous changes. The psychological stress of medical workers during the COVID-19 pandemic is discussed in [12], [13], and they were calling for an active intervention strategy to help medical workers relieve stress. A summary of the physiological metrics which can be utilized to monitor the physical health and mental well-being of COVID-19 positive individuals and frontline workers is included in [14], and they are calling for adopting wearable devices with physiological sensors to assist in alleviating the negative mental impacts brought by the pandemic. In [15], they adopted a virtual reality platform and physiological signals (EDA, ECG, PPG, and respiration impedance) to evaluate the feasibility of digital analysis and interventions to monitor and reduce frontline healthcare providers' moral distress. Not only medical workers, people working in other industries and students who are not having normal campus life are also affected by COVID-19. In [16], EDA helped evaluate the stress of people working remotely due to COVID-19. In [17], they developed an artificial electronic nose system and used EDA signals to detect the academic stress of engineering students at a university in Colombia.

The above works target users in specific groups undergoing specific activities. Our work focuses on general users of all ages undergoing a regular daily routine. We examine EDA signals collected from wrist-worn wearable devices since these devices can be worn in all scenarios to detect human stress without intruding on daily activities.

III. System Overview

The proposed framework is illustrated in Fig. 1. A wrist-worn device, such as a smartwatch, collects the EDA signals from

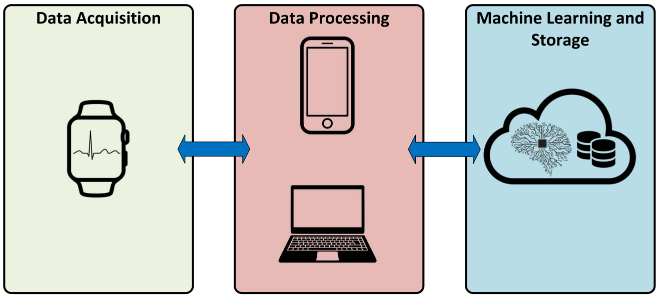


Fig. 1. The framework of the proposed stress detection system. There are three main components: i) the wearable device that collects the signal, ii) the edge devices that performs the basic signal processing and provides initial feedback, and iii) the cloud where the data is stored, while advanced machine learning algorithms that can help with further processing.

a subject. The signal is forwarded wirelessly to a computing device, such as a nearby smartphone or a computer. Then, it undergoes several signal processing steps, and the final extracted features are used for classification. Machine learning methods that can be executed in a smartphone or a device with limited computational power are used for the classification and the final binary decision of the stress status. An immediate response can go back to the wearable device. The data can be forwarded from the edge devices to the cloud if further processing is needed. When the results are available, the cloud can forward them back to the smartphone or the computer.

A. EDA Signal and Characteristics

EDA is the property of the human body that causes continuous variation in the electrical characteristics of the skin. An EDA sensor can measure the fluctuations in skin conductivity that are caused by sweat secretion. An EDA signal has two primary components: tonic and phasic levels. Tonic level and phasic level have opposite characteristics. The tonic level is a relatively stable feature in EDA, which fluctuates smoothly and slowly according to individual skin moisture level and autonomous adjustment ability. The skin conductance level (SCL) denotes the measurement of tonic level, and it is the baseline of EDA. Compared to the tonic level, the phasic level has a solid response to stimuli and is a fast-response component in EDA [18]. Skin conductance response (SCR) denotes the measurement of the phasic level. After the human body is affected by stimuli, sweat is secreted to cause changes in skin conductance, and SCR fluctuations are generated. The changes of SCR in the phasic level are more dramatic and rapid than SCL in the tonic level. The fluctuations of SCR can be observed in the form of bursts or peaks. There are event-related skin conductance responses (ER-SCRs) and non-specific skin conductance responses (NS-SCRs). Usually, one to five seconds after the stimulation, ER-SCRs will occur. On the contrary, NS-SCRs happen unconsciously or without identifiable stimuli. It is difficult to measure the SCL directly since uncontrollable NS-SCRs will exist in the EDA raw signal, even if no intended stimuli are provided to the subjects. Filters decompose and calculate the SCL and SCR components when processing raw EDA signals. The SCR's latency and amplitudes in response to stimuli are the main focus of researchers, for the signals contain information for emotional arousal to the stimuli, especially the ER-SCR, which is significantly related to a stimulus and can indicate how much the participants are engaged in the activities. Four main features, shown in Fig. 2, can be extracted and utilized in the ER-SCRs.

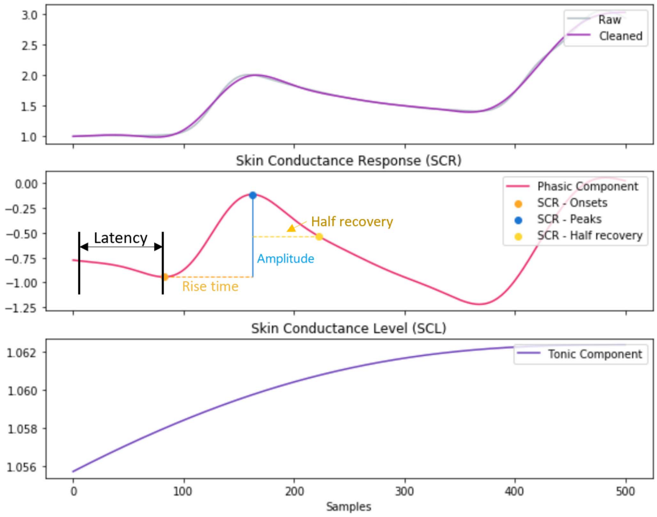


Fig. 2. The two components and four features of an EDA signal.

1) Latency. The duration from stimulus onset to the onset of the phasic burst.
2) Peak amplitude. The amplitude difference between onset and peak.
3) Rise time. The duration from onset to peak.
4) Recovery time. The duration from peak to 100% recovery.

B. Wrist-Worn Wearable Devices With EDA Sensor

Wearable devices use software and hardware components to achieve powerful functions through data and cloud interaction. In particular, these devices combine technologies such as multimedia, sensors, and wireless communication with daily wear to realize hardware terminals with functions such as user interaction, entertainment, and physiological monitoring. Medical treatment is one of the main directions of wearable devices' development.

Smartwatches are popular wearable devices that act as mini smartphones with computational and data transmission capabilities. As their popularity increases, the amount of sensor data that can collect and transmit also increases. These data can be enough to perform machine learning methods for classification and prediction tasks. The wrist-worn wearable devices that currently have EDA sensors and are available on the market are shown in Table I.

C. Publicly Available EDA Datasets

There are publicly available datasets that contain physiological signals collected from different experiments. In these experiments, signals are obtained from different skin parts according to the devices and the research objectives. For instance, the signals

TABLE I
WRIST-WORN WEARABLE DEVICES WITH EDA SENSOR AVAILABLE ON THE MARKET

Device	Sensors	Functions
Empatica E4	EDA, PPG, ST	Collect physiological data
Fitbit Sense	ECG, EDA, ST, PPG, SpO2	Stress management,monitor HR&HRV, SpO2, Breathing Rate, Sleep quality, ST, and ECG
HEALBE GoBe2	Impedance, HR, EDA	Track calorie intake, body hydration, sleep quality, heart rate, and stress levels
MOXO Sensor	EDA	Measure emotional reactions
MyFeel	HR, EDA, ST	Monitor stress level

TABLE II
AVAILABLE DATASETS WITH WRIST-BASED SIGNALS

Dataset	Subjects	Device	Activities	Modalities
CLAS [20]	62	Shimmer3	Math & logic problems, stroop test, watching videos and pictures	EDA, ECG, PPG
UTD [22]	20	Affectiva Q Curve, Nonin WristOx2	Walking, counting, stroop test, watching videos	EDA, Temp, HR SpO2, 3D Acc
VerBIO [23]	55	Empatica E4, Actiwave Cardio	Public speaking	EDA, ECG, BVP, Temp, 3D Acc
WESAD [24]	15	RespiBAN, Empatica E4	Reading, watching videos, public speaking, meditation, a mental arithmetic task	EDA, ECG, BVP, RESP, EMG, Temp, ACC

in PPG-DaLiA [19] and CLAS [20] datasets were acquired from the wrist, and the EDA signals in ITMDER [21] were recorded from fingertips or palm.

In this study, we used four publicly available datasets to develop the proposed stress detection system, including the CLAS [20], UTD [22], VerBIO [23], and WESAD [24]. The physiological signals in the four datasets are used to train and test the classification models. An overview of the datasets is available in Table II. The following are the details of the datasets.

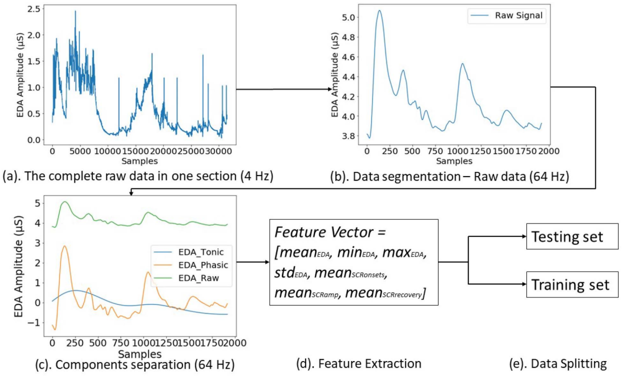
1) CLAS: The CLAS dataset is designed to study intelligent human-computer interaction (HCI). This dataset includes a group of automated human psychological and physiological evaluations, such as automatically detecting emotions and stress conditions. CLAS includes EDA, ECG, PPG, and accelerometer signals when 62 participants addressed different problems. However, three subjects' EDA data are incomplete so that 59 subjects' EDA data are used in this study. The participants were asked to work on interactive and perception tasks in this study. In the interactive task, the participants should answer mathematical and logical questions quickly to estimate the participants' cognitive load and concentration level. The authors selected images and video clips to prompt the participants' emotional arousal in the perception task. The participants answered self-assessment questions after each task, and the self-assessments are considered ground truth labels. Shimmer3 GSR+ Unit was used to collect the EDA signals in CLAS with 256 Hz. The EDA signals are collected from the fingers, though the device is placed on the wrist.

2) UTD: UTD was built to identify responses to various types of stress: cognitive stress, emotional stress, physical stress, and relaxation in this research. Twenty university students attended this research and joined the seven-stage activities. The physiological signals in this dataset were recorded by wrist-worn devices with EDA, temperature, acceleration, HR, and SpO2 sensors.

3) VERBIO: The goal of building the VerBIO dataset was to learn whether stress could influence physiological signals during public speaking. Fifty-five speakers gave 344 public speeches, and physiological signals were collected during the speeches. However, only eighteen subjects had EDA data from Empatica E4 during both PRE and POST speech sections, and the EDA data from these eighteen subjects were used for this study. In different conferences, the speakers needed to give speeches to either real or virtual audiences. The speakers delivered their speeches to virtual audiences with virtual reality equipment in virtual mode. The authors used Empatica E4 to record the EDA data with a frequency of 4 Hz and labelled the data with the speakers' self-report. The self-reports were taken before and after each session.

4) WESAD: WESAD was built to study the feasibility of recognizing emotional states with physiological signals. It includes EDA, ECG, EMG, respiration, body temperature, and triaxial acceleration data. Fifteen participants were recruited for this study and asked to watch videos, speak publicly, solve mental arithmetic problems, and meditate during the experiments. Same as VerBIO, Empatica E4 obtained the EDA data in WESAD with a frequency of 4 Hz. This research extended the authors' previous study about stress and emotion detection by introducing three emotional states: neutral, stress, and entertainment. The participants were asked to complete self-report questionnaires after each activity.

In the proposed system, there are two main reasons for using these four datasets. First, the EDA data in all four datasets were recorded by different devices. The wristband Empatica E4 was used for collecting the data in VerBIO and WESAD, while the Affectiva Q Curve was used for UTD. Both Empatica E4 and Affectiva Q Curve have electrodes built into the wristband to detect the EDA signals on the wrist [25]. In contrast, CLAS used Shimmer3, a wrist-worn device that collects the EDA signal from the fingertips. The different signal resources can provide evidence for comparing the data's reliability and quality. Second, the experiments in the four datasets have similar activities. For example, subjects in CLAS, UTD, and WESAD were asked to watch videos; math problem solving was one of CLAS, UTD, and WESAD; public speaking was required to perform in VerBIO and WESAD. The objectives of building these four



(a). The complete raw data in one section (4 Hz)

(b). Data segmentation – Raw data (64 Hz)

(c). Components separation (64 Hz)

(d). Feature Extraction

(e). Data Splitting

Fig. 3. In one section, (a) the original raw data is processed in four steps, including (b) data segmentation, (c) components separation, (d) feature extraction, and (e) data splitting.

TABLE III
SEGMENTATION OVERVIEW FOR EACH DATASET

Dataset	Selected Subjects	No. of Segments	
		Stress	Non-stress
CLAS	59	1238	603
UTD	20	451	687
VerBIO	18	284	322
WESAD	15	317	572

datasets were for emotion/stress detection. In this way, there can be more possibilities to cross-compare the data.

Furthermore, except for the CLAS, the other three datasets have detailed demographic descriptions, including the age and gender of the subjects, enabling further research on potential correlation. For instance, correlations between stress levels and demographic features such as age and ethnicity could be examined.

IV. METHODOLOGY

A stress detection system should map the inputs, which are physiological signals, to an output representing the stress level or status. We address this problem by formulating it as a binary classification task. Let $E = [e_1, e_2, \dots, e_t, \dots, e_n]$ be the input EDA data, where e_t denotes the EDA measurement in Siemens at discrete time t, the stress detection problem can be defined as follows:

$$s = C(E_w, \theta) \tag{1}$$

where $s = \{0, 1\}$ is the stress state, with $s = 1$ means stress and $s = 0$ means non-stress. $E_w \in E$ is a subset of EDA signals defined according to the segmentation window $w < n$. C is the classifier and θ is the corresponding coefficient. A supervised machine learning algorithm can learn the classifier and its corresponding coefficients.

A. Data Preprocessing

Four main steps are performed during data pre-processing as shown in Fig. 3.

1) Data Segmentation: Usually, EDA data is collected during different activities of the participants' everyday life so that the collecting period could last from minutes to hours, even days. However, the long duration of EDA data could be more convenient for analysis due to the high computational cost and sample inconsistency. As a result, EDA data should be segmented to a certain length so that the format of samples can be consistent and the computational cost can be reduced. This study segmented all data and labels by a 30-second non-overlapping sliding window for next-step processing. In UTD and WESAD, the data came from more than one activity and had four ground truth labels. For this study's binary classification objective (stress and non-stress), we merged the labels to stress and non-stress categories for each dataset according to the stress status indicated on the labels. We also excluded the data obtained from the physical activity in the UTD to improve the similarity of the four datasets and perform a fair comparison of the results. The excluded part of UTD is the data labelled "PhysicalStress" in the dataset, obtained when the participants were standing or walking/jogging on a treadmill. After applying the sliding window, 1840 samples were from CLAS, 1138 samples were from UTD, 889 samples were from WESAD, and 606 samples were from VerBIO. The segmentation results of the four datasets are shown in Table III. Only VerBIO is a relatively balanced dataset. The other three are imbalanced datasets.

2) Components Separation: Since raw EDA data contains redundant information, further data preprocessing is still necessary. At the same time, motion artifacts will exist since the EDA sensors move slightly on the skin, caused by body movements and skin moisture. Hence, extracting SCR and SCL components and applying artifact removal methods to the data are essential for later analysis. The cvxEDA model [26], based on Maximum a Posteriori (MAP), sparsity, and convex optimization, is used to decompose the SCR and SCL components. Since the cvxEDA algorithm relies on the probabilities of the parameters in the model, preprocessing, such as bandpass filtering, and postprocessing the signal, is not mandatory.

3) Feature Extraction: As training with all features in the signals would increase the computational cost, statistical features and additional SCR features were computed and extracted to form a feature vector used to train the data. Seven features are chosen to establish the feature vector ([24], [27]). The feature vector can be expressed as:

$$\begin{aligned} FeatureVector = [&mean_{EDA}, min_{EDA}, \\ &max_{EDA}, std_{EDA}, \\ &mean_{SCRonsets}, mean_{SCRamp}, \\ &mean_{SCRrecovery}] \end{aligned} \tag{2}$$

where the $mean_{EDA}, min_{EDA}, \& \, max_{EDA}, std_{EDA}$ are based on the actual EDA value in each signal window [24]. Meanwhile, the data with all features, on which no feature extraction is performed, is used to train the models as well as for comparing the classification results with the extracted feature vector.

TABLE IV
NUMBER OF SUBJECTS FOR TRAINING, VALIDATION, AND TESTING

Dataset	All	Training	Validation	Testing
CLAS	59	52	1	6
UTD	20	17	1	2
VerBIO	18	15	1	2
WESAD	15	12	1	2

4) Data Splitting: After the preprocessing, the data are split into training and testing sets based on subjects since more interference factors can be eliminated from the relationships between each individual's emotional status and EDA, so that subject-independent analyses are more advantageous than subject-dependent methods. We took 10% of the subjects as testing set (rounded up), and the rest are for training. The leave-one-subject-out method is used to validate the training. Hence, one subject is left for validation in every training round. This method saves more training time than leaving more than one subject out, as more subject combinations will be applied if independent subject impact needs to be investigated. In addition, the whole training and testing processes were completed based on cross-validation. We repeated the training and testing ten times with a random selection of the training, validation, and testing sets. All the results are the average of the ten tests. The data splitting results are shown in Table IV.

B. Classification Using Learning Systems

Five machine learning methods are used to perform the classification task. We will evaluate the results of the classification models.

1) K-Nearest Neighbor (KNN): KNN classifies by measuring the distance between different feature values. Specifically, given a training dataset, the task is to find the K instances closest to the input instance in the training dataset. The input instance will be classified according to most of the K instances belonging to which specific class. In this study, we applied a cross-validated grid-search method to determine the optimal K value, which is 3. The method can score the model with different parameters, and the parameter that offers the best performance will be considered the final one.

2) Support Vector Machine (SVM): SVM is a two-class classification model. Its principal is to find a hyperplane that satisfies the maximum interval between classes in the feature space. The learning strategy of SVM can transform into the solution of a convex quadratic programming problem. In addition, the kernel trick can be applied to extend SVM to a non-linear classifier. The commonly used kernel functions of SVM are the Gaussian kernel and the Sigmoid kernel. In this study, the Gaussian kernel was applied when using the SVM model to address the classification problem, since it is more suitable for data with complicated features by projecting the non-linear problem to another dimension, which can measure the similarity between the original features and the projected features. In this way, samples of the same kind can be better gathered together and then linearly separable.

3) Naive Bayes: Bayesian classification is a general term for classification algorithms based on Bayes' theorem. Therefore, they are collectively referred to as Bayesian classification. The Naive Bayes classification is the simplest and most common classification method in Bayesian classification. Naive Bayes is different from most other classification algorithms among all machine learning classification algorithms. Most classification algorithms, such as KNN and SVM, are discriminative methods. To directly learn the relationship between feature output Y and feature X, it adopts either a decision function $Y = f(X)$ or a conditional distribution $P(Y|X)$. However, Naive Bayes is a generation method, which means directly finding the joint distribution $P(X, Y)$ of feature output Y and feature X, and then calculating the results based on $P(Y|X) = P(X, Y)/P(X)$. In this study, Naive Bayes can provide a perspective from the assumption of independence determined by the conditional probability distribution for the classification of the EDA data.

4) Logistic Regression: Logistic Regression is a linear classification algorithm investigating a sample's probability of belonging to a particular category. Logistic Regression calculates the best decision boundary to distinguish the categories the most. Logistic Regression can be described as a discriminative model, which means the model can directly learn the decision function $Y = f(x)$ or the data's conditional probability distribution $P(Y|X)$. KNN, SVM, and Random Forest belong to discriminative models as well. In this study, Logistic Regression is applied to address the binary classification problem, as the method does not require the variables to be continuous and linear.

5) Random Forest: Random Forest is to build a forest randomly. There are many decision trees in the forest, and each tree is trained independently. When there is a new sample, each decision tree in the forest decides which category the sample belongs. Then, the decision trees vote to determine the final classification result based on which category is more selected. The Random Forest outputs the average of all decision tree outputs in the regression problem. A Random Forest can be used for both classification and regression. It is also a dimensionality reduction method that deals with missing values and outliers. Random Forest is chosen to tackle the classification problem here because this method can handle high-dimensional data and does not need to make feature selections, so that it has strong adaptability to data sets that meet the characteristics of the data in this study.

C. Evaluation Metrics

The accuracy, recall, precision, and F1-score, all of which are standard statistical evaluation methods, are used to analyze the classification performances. Accuracy can be used when the class distribution is similar, while F1-score is better for imbalanced classes. The metrics can be described as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

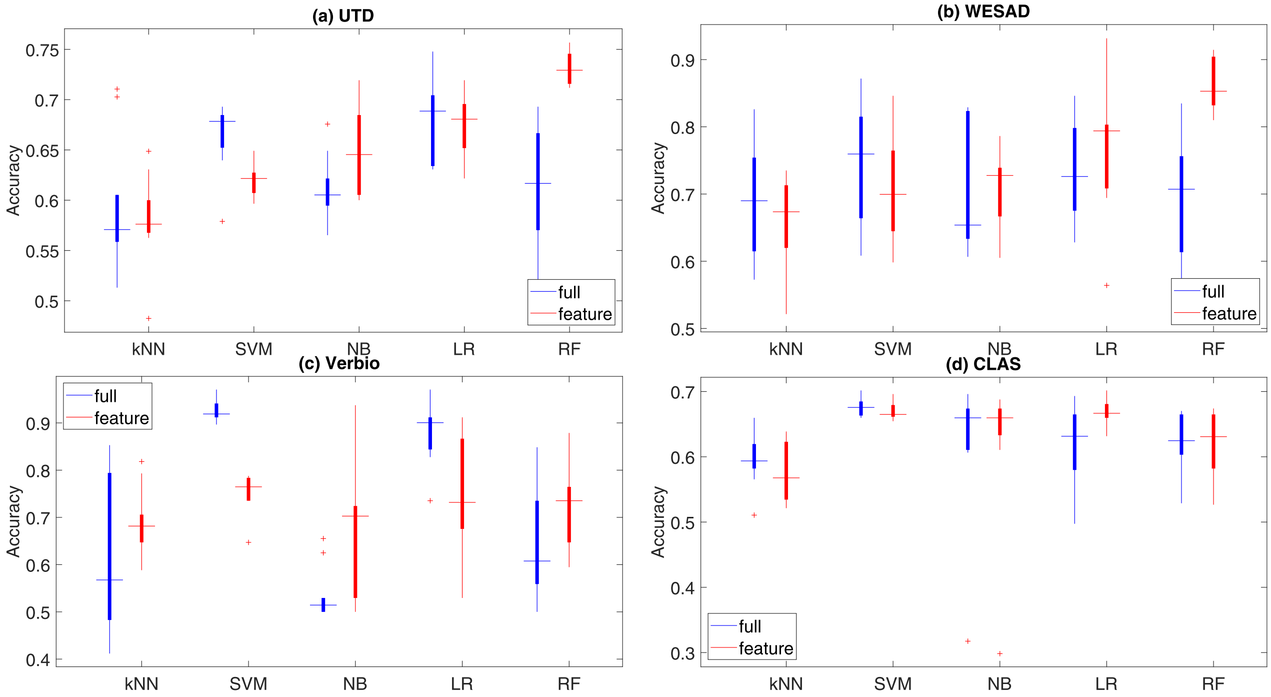


Fig. 4. Box plots of the algorithm comparison on the four datasets. In each box plot, the top and bottom whiskers represent the range of the training accuracy. The box represents the distribution of the training accuracy via the quartiles, and the horizontal bar on the box represents the median. The blue and red boxes represent the result distributions of training with all features and extracted features, respectively.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$F1 - score = 2 * \frac{precision * recall}{precision + recall} \tag{6}$$

$$TP = True\,positive, TN = True\,negative,$$

$$FP = False\,positive, FN = False\,negative.$$

V. Results and Discussion

We evaluated the performance of the proposed system for stress detection accuracy. Then, we examined the extracted features' effects on the accuracy and the performance of the different machine learning. We further investigate any impact of gender on stress detection accuracy. Finally, we explore a multimodal approach for stress detection, including two other signals, PPG and ECG, both available in wrist-worn wearable devices.

A. Evaluation of Extracted Features

We performed the binary (stress and non-stress) classification task with the data with all features and the feature vector with extracted features, respectively. As shown in Fig. 4, no consistency can be found in all four datasets regarding the differences between training with all features or extracted features. No pattern for which features will be more suitable for training each dataset is found. KNN had better performances when training with all features than extracted features on CLAS and WESAD, while it shows the contrary results on UTD and VerBIO. Logistic Regression only worked well with all features on UTD, and Random Forest only had better performance with all features on CLAS. However, it should be mentioned that SVM and Naive Bayes show a pattern that had better results when training with one set of features over the other one. SVM prefers all features, while Naive Bayes is the opposite.

The results also showed that the Naive Bayes model has the lowest theoretical error rate compared with other classification methods. However, this is only sometimes the case since Naive Bayes assumes that the features are independent when given output categories. This assumption is often invalid in practical applications. When the correlation of features is significant, the classification effect is less than when the correlation is small. In the EDA data, the extracted features are relatively independent of boosting the Naive Bayes' performance. The essence of the SVM learning strategy is to maximize the interval between categories, thereby turning the classification problem into a problem of solving convex quadratic programming. At the same time, the method of SVM to solve nonlinear problems is to transform them into linear classification problems in another feature space through nonlinear transformation. These principles lead to the decision boundary needing improvement, determined by a few features or support vectors. Though the other three models do not show consistency about which features are preferable, there is still a tendency that extracted features have an overall advantage for training the models as high dimensions will cause the over-fitting problem.

Following these insights, it should be clear that it is unreliable to determine whether training with all or extracted features is safe for different datasets. Consequently, comparing training

TABLE V
EVALUATION OF THE CLASSIFIERS ON STRESS DETECTION WHEN ALL FEATURES AND EXTRACTED FEATURES ARE USED

Model	Data	Metrics	Dataset			
			CLAS	UTD	VerBIO	WESAD
KNN	All features	F1-score	0.714	0.336	0.613	0.554
		Accuracy	0.595	0.569	0.620	0.691
	Extracted features	F11score	0.696	0.351	0.669	0.476
		Accuracy	0.576	0.586	0.691	0.661
SVM	All features	F1-score	**0.801**	0.292	**0.879**	0.614
		Accuracy	**0.685**	0.663	**0.929**	0.752
	Extracted features	F1-score	0.798	0.077	0.730	0.385
		Accuracy	0.664	0.62	0.754	0.714
Naive Bayes	All features	F1-score	0.744	0.388	N/A	0.596
		Accuracy	0.635	0.586	0.538	0.678
	Extracted features	F1-score	0.73	0.299	0.602	0.447
		Accuracy	0.632	0.648	0.679	0.712
Logistic Regression	All features	F1-score	0.746	0.380	0.840	0.407
		Accuracy	0.639	0.68	0.877	0.731
	Extracted features	F1-score	0.799	0.353	0.621	0.637
		Accuracy	0.666	0.676	0.732	0.781
Random Forest	All features	F1-score	0.727	0.394	0.643	0.561
		Accuracy	0.619	0.613	0.64	0.683
	Extracted features	F1-score	0.735	**0.582**	0.714	**0.792**
		Accuracy	0.615	**0.731**	0.716	**0.865**

with all features and different combinations of extracted features is necessary.

B. Evaluation of ML Classification

The performances of each machine learning model are shown in Table V. According to the results, the accuracy of Random Forest is 73.1% and 86.5% for UTD and WESAD, respectively, while SVM can reach a classification accuracy of 92.9% for VerBIO. For CLAS, SVM has better performance than the other four methods. KNN, Naive Bayes, and Logistic Regression cannot offer satisfactory results for any datasets. Although KNN supports non-linear solutions, the co-linearity and outliers in the data are expected to be processed before training. Naive Bayes expects all features to be independent, and Logistic Regression provides linear solutions and assumes that input features have no co-linearity. On the other hand, SVM and Random Forest can efficiently support non-linear data and cope with co-linearity. For this study, the EDA samples in the datasets have close correlations since the SCR features (peak amplitude, onsets, recovery, etc.) are physiological responses, and the various physical and emotional changes of human beings are influenced and interact with each other. Consequently, the performances of SVM and Random Forest outweigh the other methods.

Additionally, for VerBIO, F1-score and accuracy are close, whether training with all features or extracted features. However, this is different for CLAS, UTD, and WESAD. The possible reason for this is that the distribution of stress and non-stress data in VerBIO is balanced, while the data distribution of CLAS, UTD, and WESAD is not balanced, as presented in Table III. Especially in CLAS, the sample distribution is highly biased, resulting in the F1-scores being much higher than the accuracy of every method. In the case of CLAS, the F1-score is more informative than accuracy.

Another interesting insight is that, overall, CLAS has the lowest accuracy of 68.5%, while with any of the methods, the accuracy in CLAS is always among the lowest. It is important to mention that the EDA signals in CLAS are collected from the fingers. This might indicate that extra noises, probably motion artifacts, are added to the signal, or information essential to EDA needs to be better captured from the finger when compared with wrist-based EDA signal, since wrists would be steadier for sensors to collect data than fingers.

A comparison between other research works that also perform binary classification (stress and non-stress) solely with EDA signals from the same datasets used in this study, is shown in Table VI. For CLAS and WESAD, our methods and results are promising compared to other studies. There is no comparison with VerBIO and UTD, since the research on these datasets focuses on different topics or uses other evaluation metrics.

C. Evaluation of the Impact of Gender

We further investigate whether other factors, such as gender, would influence the classification results. We trained the models separately with male and female subjects in UTD, VerBIO, and

TABLE VI
A Comparison Between Similar Works Based on the Same Datasets

Dataset	Model	Acc.	Features	Ref.
VerBIO	SVM	**0.929**	All features	This study
UTD	Random Forest	**0.731**	Extracted features	This study
CLAS	CNN	0.664	Extracted features	[28]
CLAS	SVM	**0.685**	All features	This study
WESAD	Random Forest	0.783	Extracted features	[29]
WESAD	AdaBoost DT	0.797	Extracted features	[24]
WESAD	Random Forest	0.842	Extracted features	[30]
WESAD	Random Forest	**0.865**	Extracted features	This study

TABLE VII
Stress Accuracy When Subject Gender is Used

Dataset	**All Subjects**	**Male Subjects**	**Female Subjects**
UTD	0.731	0.676	**0.716**
VerBIO	0.929	0.716	**0.876**
WESAD	0.865	0.656	**0.787**

TABLE VIII
Detection Accuracy for Different Modalities of CLAS

CLAS	**Classification**				
	SVM	KNN	RF	NB	LR
EDA	**0.699**	**0.688**	0.645	**0.634**	0.651
ECG	0.677	0.603	0.602	0.423	0.559
PPG	0.645	0.581	0.629	0.597	0.629
EDA+PPG	0.683	0.667	**0.694**	0.29	**0.704**
EDA+ECG	0.661	0.538	0.624	0.296	0.661
PPG+ECG	0.688	0.613	0.677	0.323	0.634
EDA+ECG+PPG	0.667	0.570	0.618	0.339	0.559

TABLE IX
Detection Accuracy for Different Modalities of VerBIO

VerBIO	**Classification**				
	SVM	KNN	RF	NB	LR
EDA	**0.929**	**0.62**	0.64	0.538	**0.877**
ECG	0.612	0.493	0.806	0.627	0.672
PPG	0.582	0.567	0.507	0.537	0.552
EDA+PPG	0.683	0.667	0.694	0.29	0.704
EDA+ECG	0.716	0.612	**0.896**	**0.687**	0.821
PPG+ECG	0.567	0.458	0.761	0.537	0.507
EDA+ECG+PPG	0.582	0.567	0.836	0.671	0.672

WESAD, and we excluded CLAS since it does not provide gender information. In UTD, 14 out of 20 subjects are males, and the other 6 subjects are females. In VerBIO, the numbers of both male and female subjects equal 9. In WESAD, 12 out of 15 subjects are males, and 3 subjects are females. Gender-specific models are trained and tested on the data from each gender only.

As shown in Table VII , the overall training results with female subjects are higher than male subjects on all three datasets. For simplicity, we included only the highest accuracy results for each method. For UTD, the best results are with RF, VerBIO with NB, and WESAD with LR. Males and females usually have different mental and physical responses to stress [31], and gender-based models can better capture the other responses. Due to the differences between the brain activities and hormonal changes of males and females, they can have different reactions to the same stress stimuli [32]. When males and females react differently to stress, females tend to generate greater tonic and phasic EDA signals [33]. This affects the EDA signals of females more than males so that the samples from females can reflect the stress status better. Finally, since the ground truth labels for all the datasets came from self-report, this might also be an insight that females could express their emotional changes more accurately in such surveys.

D. Evaluation of the Effect of Multimodal Fusion

We further trained the classification models with multiple modalities available in a wrist-worn wearable device and evaluated different physiological signals' potential contribution to stress detection. We focus on ECG and PPG to detect people's emotional changes as well as EDA since these sensors are available in most smartwatches and the data are available in three of the four datasets. ECG can record the timing and different electrical discharges associated with heartbeats. PPG is a non-invasive detection method that detects blood volume changes in living tissue by photoelectric. When people encounter a stimulus, both heartbeats and vasoconstriction could affect ECG and PPG. As a result, adopting ECG and PPG signals can be informative in performing stress detection.

Multimodal signals are usually fused at three levels: sensor-level fusion, feature-level fusion, and decision-level fusion [34]. In this work, the three modalities from different sensors are concatenated to form a new feature vector before being fed into the classifiers. The signal fusion was realized at the feature level. Extracted features of ECG and PPG are used to establish the new feature vector. For ECG signals, principal component analysis (PCA) extracted the most relevant features from the heart rate (HR) and heart rate variability (HRV) time series of the ECG signals. HR and HRV are calculated from the Inter Beat Intervals (IBI), which are acquired based on the R-peaks of the ECG signals. For PPG, the maximum heart rate after stimulus onset (PPG_Rate_Max), the minimum heart rate after stimulus onset (PPG_Rate_Min), the mean heart rate after stimulus onset (PPG_Rate_Mean), and the standard deviation of the heart rate after stimulus onset (PPG_Rate_SD) are the features used in this study.

Since UTD does not contain PPG and ECG signals, only CLAS, VerBIO, and WESAD are examined. Tables VIII, IX, and X, show the detection accuracy based on different modalities as well as different combinations among them, for CLAS,

TABLE X
DETECTION ACCURACY FOR DIFFERENT MODALITIES OF WESAD

WESAD	**Classification**				
	SVM	KNN	RF	NB	LR
EDA	**0.839**	**0.754**	0.763	0.746	**0.737**
ECG	0.644	0.381	0.669	0.559	0.517
PPG	0.678	0.602	0.788	0.644	0.559
EDA+PPG	0.686	0.449	0.619	0.636	0.508
EDA+ECG	0.636	0.534	**0.822**	0.788	0.552
PPG+ECG	0.627	0.458	0.805	**0.797**	0.576
EDA+ECG+PPG	0.593	0.424	0.636	0.593	0.508

VerBIO, and WESAD dataset, respectively. EDA outperforms ECG and PPG accuracy, as well as signal combinations for stress detection. It should be mentioned that only in Table X, the highest detection accuracy of Naive Bayes was offered by the PPG and ECG fusion, which is non-relevant with EDA. All the other best performance of each machine learning model is either based on a single EDA signal or multimodal, including EDA. The overall optimal result of 92.9% is still from the SVM model based on EDA from the VerBIO dataset. This can be an indicator that EDA can directly reflect people's emotional changes. Moreover, the fluctuations of ECG and PPG may not be as sensitive to minor emotional changes as SCR. As a result, EDA should be the primary candidate when performing emotion-related detection. Additionally, for nonstationary physiological signals, the frequency-domain features may offer better discrimination ability of the physiological responses than the time-domain features since the spectral information in the frequency domain can represent oscillation information. When the classification efficiency of adopting time-domain features in PPG and ECG is unsatisfactory, extracting and analyzing frequency-domain features and performing time-frequency analysis should be considered to evaluate the signals' significance in specific tasks [35], [36].

VI. Conclusion

This work evaluates the classification performances of five machine learning models on four EDA datasets. We trained the models with all features and extracted EDA and SCR features separately. The results showed that Random Forest offered the best binary classification performances on UTD and WESAD, the accuracy of which is 73.1% and 86.5%, and SVM reached an accuracy of 92.9% for VerBIO. Additionally, this study used ECG, PPG, and their fused multimodal with EDA to evaluate the influence of different modalities for stress detection. The results show that EDA outweighs the other modalities. An interesting finding from this study is that there is a significantly better relation between EDA and stress of female subjects than male subjects. The proposed framework and the experimental results show the feasibility of using wrist-worn wearable devices with EDA sensors for stress detection and classification.

References

[1] L. Holtz, M. Martinez, K. Paton, K. Rosich, and E. Schnittka, "Effects of physiological stress response on short-term memory recall," *J. Adv. Student Sci.*, 2017.

[2] J. Ralls, "Fitbit debuts sense, its most advanced health smartwatch; World's first with EDA sensor for stress management, plus ECG APP, SPO2 and skin temperature sensors," 2021. [Online]. Available: https://www.businesswire.com/news/home/20200825005373/en/

[3] P. Sariñana-González, Á. Romero-Martínez, and L. Moya-Albiol, "Cooperation induces an increase in emotional response, as measured by electrodermal activity and mood," *Curr. Psychol.*, vol. 36, no. 2, pp. 366–375, 2017.

[4] C. Marzi, A. Greco, E. P. Scilingo, and N. Vanello, "Towards a model of arousal change after affective word pronunciation based on electrodermal activity and speech analysis," *Biomed. Signal Process. Control*, vol. 67, 2021, Art. no. 102517.

[5] A. Szpak, T. Loetscher, O. Churches, N. A. Thomas, C. J. Spence, and M. E. Nicholls, "Keeping your distance: Attentional withdrawal in individuals who show physiological signs of social discomfort," *Neuropsychologia*, vol. 70, pp. 462–467, 2015.

[6] H. F. Posada-Quintero, J. P. Florian, A. D. Orjuela-Cañón, and K. H. Chon, "Electrodermal activity is sensitive to cognitive stress under water," *Front. Physiol.*, vol. 8, 2018, Art. no. 1128.

[7] B. Choi, H. Jebelli, and S. Lee, "Feasibility analysis of electrodermal activity (EDA) acquired from wearable sensors to assess construction workers' perceived risk," *Saf. Sci.*, vol. 115, pp. 110–120, 2019.

[8] A. Anusha et al., "Electrodermal activity based pre-surgery stress detection using a wrist wearable," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 1, pp. 92–100, Jan. 2020.

[9] P. Zontone, A. Affanni, R. Bernardini, L. Del Linz, A. Piras, and R. Rinaldo, "Emotional response analysis using electrodermal activity, electrocardiogram and eye tracking signals in drivers with various car setups," in *Proc. 28th Eur. Signal Process. Conf.*, 2021, pp. 1160–1164.

[10] Y. Liu and S. Du, "Psychological stress level detection based on electrodermal activity," *Behav. Brain Res.*, vol. 341, pp. 50–53, 2018.

[11] A. Mestanikova, I. Ondrejka, M. Mestanik, I. Hrtanek, E. Snircova, and I. Tonhajzerova, "Electrodermal activity in adolescent depression," in *Pulmonary Infection and Inflammation*. Berlin, Germany: Springer, 2016, pp. 83–88.

[12] X. Shen, X. Zou, X. Zhong, J. Yan, and L. Li, "Psychological stress of ICU nurses in the time of COVID-19," *Critical Care*, vol. 24, pp. 1–3, May 2020.

[13] W. Wu et al., "Psychological stress of medical staffs during outbreak of COVID-19 and adjustment strategy," *J. Med. Virology*, vol. 92, no. 10, pp. 1962–1970, 2020.

[14] D. R. Seshadri et al., "Wearable sensors for COVID-19: A call to action to harness our digital infrastructure for remote patient monitoring and virtual assessments," *Front. Digit. Health*, vol. 2,2020, Art. no. 8.

[15] B. Nguyen et al., "Digital interventions to reduce distress among health care providers at the frontline: Protocol for a feasibility trial," *JMIR Res. Protoc.*, vol. 11, no. 2, 2022, Art. no. e32240.

[16] M. Virtaneva et al., "COVID-19 remote work: Body stress, self-efficacy, teamwork, and perceived productivity of knowledge workers," in *Proc. 12th Scand. Conf. Inf. Syst. Assoc. Inf. Syst.*, 2021.

[17] C. M. D. Acevedo, J. K. C. Gómez, and C. A. A. Rojas, "Academic stress detection on university students during COVID-19 outbreak by using an electronic nose and the galvanic skin response," *Biomed. Signal Process. Control*, vol. 68, 2021, Art. no. 102756.

[18] J. J. Braithwaite, D. G. Watson, R. Jones, and M. Rowe, "A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRS) for psychological experiments," *Psychophysiology*, vol. 49, no. 1, pp. 1017–1034, 2013.

[19] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, "Deep PPG: Large-scale heart rate estimation with convolutional neural networks," *Sensors*, vol. 19, no. 14, 2019, Art. no. 3079.

[20] V. Markova, T. Ganchev, and K. Kalinkov, "CLAS: A database for cognitive load, affect and stress recognition," in *Proc. IEEE Int. Conf. Biomed. Innovations Appl.*, 2019, pp. 1–4.

[21] J. Pinto, "Exploring physiological multimodality for emotional assessment." Lisboa, Portugal: Instituto Superior Tcnico, 2019.

[22] J. Birjandtalab, D. Cogan, M. B. Pouyan, and M. Nourani, "A non-EEG biosignals dataset for assessment and visualization of neurological status," in *Proc. IEEE Int. Workshop Signal Process. Syst.*, 2016, pp. 110–114.

[23] M. Yadav, M. N. Sakib, E. H. Nirjhar, K. Feng, A. Behzadan, and T. Chaspari, "Exploring individual differences of public speaking anxiety in real-life and virtual presentations," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1168–1182, Jul.–Sep. 2022.

[24] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proc. 20th ACM Int. Conf. Multimodal Interaction*, 2018, pp. 400–408.

[25] Empatica, "E4 wristband," 2020. Accessed: Jul. 13, 2021. [Online]. Available: https://www.empatica.com/en-int/research/e4/

[26] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxEDA: A convex optimization approach to electrodermal activity processing," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 797–804, Apr. 2016.

[27] P. Bobade and M. Vani, "Stress detection with machine learning and deep learning using multimodal physiological data," in *Proc. 2nd Int. Conf. Inventive Res. Comput. Appl.*, 2020, pp. 51–57.

[28] R. K. Radhika and V. R. Murthy Oruganti, "Deep multimodal fusion for subject-independent stress detection," in *Proc. 11th Int. Conf. Cloud Comput., Data Sci. Eng.*, 2021, pp. 105–109.

[29] P. Siirtola, "Continuous stress detection using the sensors of commercial smartwatch," in *Proc. Adjunct ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Proc. ACM Int. Symp. Wearable Comput.*, 2019, pp. 1198–1201.

[30] P. Garg, J. Santhosh, A. Dengel, and S. Ishimaru, "Stress detection by machine learning and wearable sensors," in *Proc. 26th Int. Conf. Intell. User Interfaces-Companion*, 2021, pp. 43–45.

[31] R. Verma, Y. P. S. Balhara, and C. S. Gupta, "Gender differences in stress response: Role of developmental and biological determinants," *Ind. Psychiatry J.*, vol. 20, no. 1, pp. 4–10, 2011.

[32] J. M. Goldstein, M. Jerram, B. Abbs, S. Whitfield-Gabrieli, and N. Makris, "Sex differences in stress response circuitry activation dependent on female hormonal cycle," *J. Neurosci.*, vol. 30, no. 2, pp. 431–438, 2010.

[33] D. S. Bari, "Gender differences in tonic and phasic electrodermal activity components," *Sci. J. Univ. Zakho*, vol. 8, no. 1, pp. 29–33, 2020.

[34] R. Sharma, V. I. Pavlović, and T. S. Huang, "Toward multimodal human–computer interface," in *Advances in Image Processing and Understanding: A Festschrift for Thomas S. Huang*. Singapore: World Sci., 2002, pp. 349–365.

[35] L. G. Tereshchenko and M. E. Josephson, "Frequency content and characteristics of ventricular conduction," *J. Electrocardiol.*, vol. 48, no. 6, pp. 933–937, 2015.

[36] S. Elzeiny and M. Qaraqe, "Stress classification using photoplethysmogram-based spatial and frequency domain images," *Sensors*, vol. 20, no. 18, 2020, Art. no. 5312.