Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory Stress Detection With Smartwatches

YEKTA SAID CAN[1], **NIAZ CHALABIANLOO**[1], **DENIZ EKIZ**[1],
JAVIER FERNANDEZ-ALVAREZ[2], **GIUSEPPE RIVA**[2], **AND**
CEM ERSOY[1], (Senior Member, IEEE)

[1]Department of Computer Engineering, Boğaziçi University, 34342 Istanbul, Turkey
[2]Psychology Department, Universita Cattolica del Sacre Cuore, 20123 Milan, Italy

Corresponding author: Yekta Said Can (yekta.can@boun.edu.tr)

ABSTRACT Researchers strive hard to develop effective ways to detect and cope with enduring high-level daily stress as early as possible to prevent serious health consequences. Although research has traditionally been conducted in laboratory settings, a set of new studies have recently begun to be conducted in ecological environments with unobtrusive wearable devices. Since patterns of stress are ideographic, person-independent models have generally lower accuracies. On the contrary, person-specific models have higher accuracies but they require a long-term data collection period. In this study, we developed a hybrid approach of personal level stress clustering by using baseline stress self-reports to increase the success of person-independent models without requiring a substantial amount of personal data. We further added decision level smoothing to our unobtrusive smartwatch based stress level differentiation system to increase the performance by correcting false labels assigned by the machine learning algorithm. In order to test and evaluate our system, we collected physiological data from 32 participants of a summer school with wrist-worn unobtrusive wearable devices. This event is comprised of baseline, lecture, exam and recovery sessions. In the recovery session, a stress management method was applied to alleviate the stress of the participants. The perceived stress in the form of NASA-TLX questionnaires collected from the users as self-reports and physiological stress levels extracted using wearable sensors are examined separately. By using our system, we were able to differentiate the 3-levels of stress successfully. We further substantially increase our performance by personal stress level clustering and by applying high-level accuracy calculation and decision level smoothing methods. We also demonstrated the success of the stress reduction methods by analyzing physiological signals and self-reports.

INDEX TERMS Stress recognition, machine learning, wearable sensors, smart-phone, smartwatch, photoplethysmography, daily life physiological data.

I. INTRODUCTION

Smart-sensing, pervasive and ubiquitous technologies have become more accessible during the last decade. A variety of smart sensing devices are emerging in the market,

The associate editor coordinating the review of this manuscript and approving it for publication was Yue Zhang.

with new sensing technologies offering more personal health monitoring options. Sophisticated sensor systems can be found in most modern smartphones and smartwatches. An individual's daily routines, fitness and physical activities can be deduced using the data coming from these sensing units. This information may help individuals to better understand and adapt their behaviors to their benefit. The new
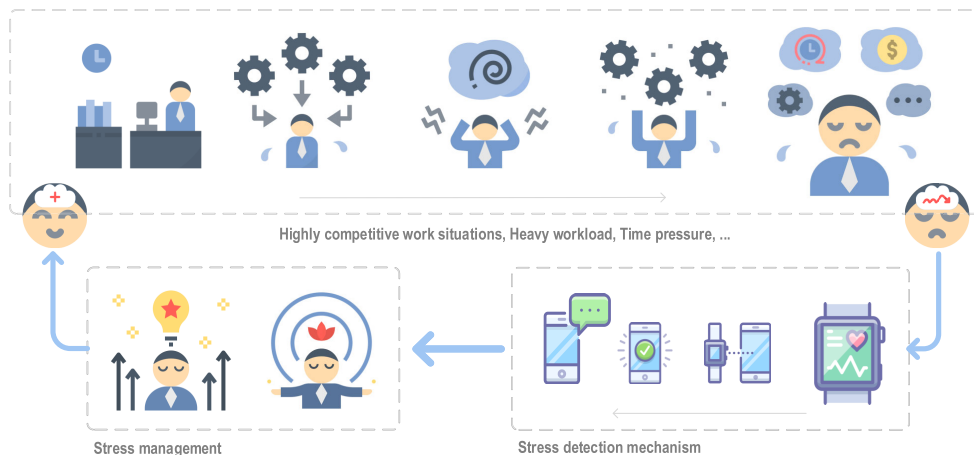
Y. S. Can *et al.*: Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory

IEEE *Access*

FIGURE 1. A case study of real-life application of our stress detection system. When a subject experiences stress, our automatic stress detection system will warn the user and offer a mobile breathing relaxation method. After the application of this technique, user will return to the baseline state.

generation of unobtrusive smart devices may be used for monitoring and improving well-being and potentially psychological health.

Stress is an important problem in the modern world. It has significant effects on human health, society and economy [1]. We encounter stressful situations in working environments, traffic, social interactions since it is impossible to avoid them. Scientists noticed that high-level and enduring stress should be managed when the first symptoms emerged in order to refrain from long-term outcomes [2]. By detecting and handling stress in the early stages, researchers and clinicians can foster the development of novel ways to tackle the problematic consequences of a sustained stress response. Until the last decade, we used to try to alleviate enduring stress by seeing experts. After the technology has taken place in our daily lives, researchers come up with ways to use it for pervasive health. The first candidates for these widespread and unobtrusive technologies are smartphones, smartwatches and smart bands which could be used comfortably in daily lives. Especially wrist-worn smart devices can collect physiological signals such as HRV (Heart Rate Variability), EDA (Electrodermal Activity), ST (Skin Temperature) which enable us for automatic stress level detection. However, they are not without their challenges which are relatively short battery lives, unable to collect data in intensive physical activity and when they are worn loosely, artifacts caused by these movements and loosely worn devices [5]. In order to overcome these problems, advanced modality specific filtering techniques should be applied.

Stress detection studies have started in the laboratory environments and then the direction of this research shifted towards real-life ambulatory environments. However, most of these studies only distinguish between stress and relax states, which is not representative of the vast array of possible nuances that define the continuum from relaxation to the highest stress levels [6]–[9]. Furthermore, cognitive load state

also frequently occurs especially in work environments and it should be added as a different class from stress to increase the stress detection resolution for more realistic results. However, when this state is added, the accuracies of stress level detection systems decrease drastically especially in real-life conditions. An important research issue is that since everyone has specific stress responses, person-independent models generally have low accuracies for finding stress levels. If the data of individuals are used for creating models, it might be insufficient or requires long-term data collection. A hybrid approach for clustering people by using their stress responses might solve the low accuracy problem without requiring long term data collection from individuals. Furthermore, after the machine learning algorithms assign labels to frames, a decision level smoothing technique could increase the performance of stress detection systems. Since the accuracies of the ambulatory systems are relatively low when compared to the systems tested in laboratory environments, they could take advantage of an extra layer of logic to fix the erroneous labels. As an example, for 1-minute frames, it is not logical for a person to have a stressed frame followed by a relaxed state and another stressed frame after that. The middle frame, in this case, could be regarded as an error and smoothed.

After detecting high levels of enduring stress, stress detection systems should also decrease the stress of individuals to acceptable levels. Stress management mechanisms should be applied to achieve this recovery (see Figure 1 for a case study of a stress detection system with relaxation method). Traditional yoga and meditation are ancient methods for stress alleviation. However, they either require outdoor environments or could not be employed in office environments during daily work routines. As a consequence, mobile apps or indoor techniques should be applied to manage stress in the daily life without interrupting daily routines. The benefit of these indoor techniques are not examined comprehensively but a limited number of studies have commenced [3]. A close

IEEE Access

Y. S. Can *et al.*: Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory

examination of the literature permits to identify that most of the stress recognition studies do not offer relaxation methods after detecting stress which is needed to recover the subjects to their baseline states.

In this work, we improved our multi-level stress detection system which uses unobtrusive smart wearable devices. To the best of our knowledge, this study is the first one to propose personal stress level clustering and decision-level smoothing to enhance person-independent stress detection models by using smartwatches. The advantage of our system when compared to systems using medical-grade devices is the applicability in the daily lives of users. Wearing devices with cables, electrodes makes people uncomfortable. However, unobtrusive devices such as smartwatches, smart bands could be worn in everyday life. The disadvantage of these devices is low data quality when compared to medical grade devices. With the application of modality-specific artifact detection and removal methods, personal stress level clustering and decision-level smoothing, our system has comparable performance with the systems developed with medical-grade devices. The other differences of our study, when compared to most of the research in the literature, are the addition of a cognitive load state besides stress and baseline states and application of a stress management method. The former improved the precision of our system with a frequently experienced real-life state and the latter increased the applicability of our system. Only detecting stress and notifying the users might increase the stress level of individuals even more. However, if the system detects the high level of stress and suggests easy to apply relaxation methods, then it could be beneficial for people for coping with stress. We tested the effectiveness of a relaxation method and its separability from the stress state with our system by using physiological signals. Our method starts with preprocessing the signals to clear the artifacts caused by unconstrained real-life motions. We further extracted discriminative features from the selected physiological signals. Lastly, machine learning algorithms are applied to classify different stress levels. To test our algorithms, we induced stress on teachers from the ILKYAR summer school seminars took place at Bogazici University. We induced stress on primary school teachers following a cognitive load session in this event and then apply a stress management technique to recover them to their baseline states. Baseline signals and self-reports from these teachers are also collected at the beginning of this event. Our decision level smoothing and stress level clustering methods are applied after the classification algorithms are applied. Guided mindfulness was further used to alleviate the stress levels. Mindfulness is a mental state achieved by focusing one's awareness on the present moment, while calmly acknowledging and accepting one's feelings, thoughts, and bodily sensations, used as a therapeutic technique in stress management. Our research addresses four original research issues:

- The effect of applying decision level smoothing and decision-making mechanisms on system performance.

- The performance evaluation of person specific, clustered according to the baseline stress levels (hybrid) and person-independent models.
- The effect of different ground truth surveys (NASA-TLX and a more suitable, less time consuming, version of it for everyday stress detection) on classification accuracies,
- Application of a guided mindfulness technique and measuring its success with smartwatch based physiological signals for reducing stress levels.

The organization of the rest of the paper is as follows. In Section 2, we present the related work in stress detection and alleviation. Our smartwatch based system for stress level monitoring is described in Section 3. ILKYAR Summer School Event and data collection are mentioned in Section 4. Experimental results and discussion are presented in Section 5. We present the conclusion of the study in Section 6.

II. RELATED WORK

Automatic stress detection research has started in laboratory environments. The stress and relax states are discriminated with high success. After researchers realized that laboratory stress is not comparable with real life stress [4], they have started working for stress level detection in real life environments. Moreover, the ultimate aim of laboratory studies is to help individuals manage their stress levels in their daily life routines. These experiments gave researchers clues about how to build such systems. However, there are more issues to deal with when the research took a step outside the laboratory e.g unrestricted movement of subjects, unknown context, reliability issue of self-reports, battery life.

The research then directed towards detecting stress levels in restricted environments such as offices, automobiles and classrooms. Traffic jams in crowded cities, offices and workplaces and exams and courses in campus environments are among the primary causes for increased stress levels. These environments could be categorized into restricted environments because the movements are limited and they can be controlled and monitored with sensors and cameras. Impactful studies in the field of office environment stress detection can be counted as [30] and [31] respectively. These studies used as EDA (Electrodermal Activity), ECG (Electrocardiography) and Accelerometer sensors respectively. For automobile environments, most of the works applied different processing and machine learning methods on DriveDB database [32] collected on Boston from 24 drivers. The database is composed of ECG, EDA, EMG (Electromyography) and respiratory sensor data. Student stress is also measured in campus environments. However, since outside classroom environments are not controlled as the mentioned restricted environments, campus stress detection studies could be regarded as a bridge from restricted environments to unlimited daily life environments [33]–[35]. There are a number of studies in unlimited daily life stress

Y. S. Can *et al.*: Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory

IEEE *Access*

TABLE 1. Recent stress detection experiments in different types of environments.

Article	Stress Signal	Method	# of Classes	Accuracy %	Environment	Relaxation Method Application
[5] (2019)	HRV, EDA and ACC	MLP, Random Forest, SVM, kNN and LDA	3 (Stressed, Cognitive Load and Baseline)	92.15	Real Life Algorithmic Contest	Yes
[6] (2018)	Usage Data for different application categories	HMM with MPM	2 (Stressed, Baseline)	68	Daily Life	No
[7] (2018)	HRV	SVM	2 (Rest (R), Stress(S))	70	Laboratory	No
[8](2018)	BioRadar	Multilayer Perceptron	2 (Rest (R), Stress(S))	0.94	Laboratory	No
[9] (2017)	Facial Cues	kNN, SVM, Naive Bayes	2 (Baseline, Stressed)	91.68	Laboratory	No
[10] (2017)	EDA, PPG, Speech, Accelerometer	Adaboost	2 (Stressed, Baseline)	94	Laboratory	No
[11] (2017)	HRV, EDA	F-state Machine	3 (Low, High stress, High alert)	0.984 (f-measure)	Laboratory	No
[12](2017)	EDA, PPG	SVM, Logistic Regression, Random Forest	2 (S, R)	0.79 (f-measure)	Laboratory	Yes
[13] (2016)	Mobile Application Usage Pattern, Physical Activity, Light Sensor, Screen Events	SVM, ANN, kNN	2 (Stressed, Baseline)	70	Daily Life	No
[14] (2016)	BVP, SkinTemperature, EDA, RR, HeartRate (Without Context Info)	RandomForest	2 (Stressed, Baseline)	76 (With Context Information 92)	Daily Life	No
[15] (2015)	Speech, Gyroscope, Accelerometer, Light Sensor, Screen Events, Activity Type	RandomForest, Simple Logic, DecisionTree	2 (Stressed, Baseline)	77	Daily Life	No
[16] (2014)	ECG, SkinTemperature, Respiration, Accelerometer, EDA	SVM, kNN, ANN, RandomForest	2 (Stressed, Baseline)	73	Daily Life	No
[17] (2013)	Call, SMS Statistics, GPS, Screen On/Off, Accelerometer, EDA	SVM, SVM-RBF, kNN	2 (Stressed, Baseline)	75	Daily Life	No
[18] (2015)	ECG + Respiratory + Accelerometer	SVM	2 (Stressed, Baseline)	72	Daily Life	No
[19] (2017)	HR, IBI, HRV, EDA, Temperature	Weka Toolkit	2 (Stressed, Baseline)	70 (precision with 95% recall)	Daily Life	No
[20] (2010)	EDA, PPG	kNN, FDA	2 (S,R)	95	Laboratory	No
[21] (2015)	EDA, PPG	SVM	2 (S,R)	80	Laboratory	No
[22] (2016)	EEG	SVM	4 (Neutral, Medium, Low, High Stress)	89	Laboratory	No
[23] (2015)	Body Movements	SVM	2 (S, R)	77	Laboratory	No
[24](2016)	Body Movements, EMG, EDA, Respiration	SVM	2 (Stressed, Baseline)	85	Laboratory	No
[25] (2012)	Temperature, Heat Flux, EDA, Respiration, Accelerometer	Naive Bayes	2 (Stressed, Baseline)	82	Laboratory	No
[26] (2017)	ECG, GSR, respiration, Blood Pressure, Blood Oximeter	SVM, kNN	2 (Stressed, Baseline)	95.8	Laboratory	Yes
[27](2014)	EEG, ECG, EMG, EOG	ANN	3 (Baseline, Mental, Fatigue Stress)	80	Laboratory	No
[28] (2015)	Facial Blood Flow	Multiple Regression	2 (S, R)	88.6	Laboratory	No
[29] (2015)	EDA	LDA	2 (S, R)	98.88	Laboratory	No

level detection with smartphones and smart wearables [6], [14], [19]. Due to the above-mentioned reasons, their accuracies are lower than restricted environments. Classification accuracies for 2-class is around 70% in daily life environments [13] (2016), [15] (2015), [19] (2017) and [6] (2018) (see Table 1). One reason for these low accuracies might be

IEEE *Access*

Y. S. Can *et al.*: Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory

the uniqueness of the stress responses of individuals. To overcome this problem, researchers developed person-specific models by only using the data from the individuals. However, this approach requires a long data collection period. We need a hybrid approach which has higher accuracies than person-independent models. If people with similar stress responses could be clustered in a hybrid approach, ML models for these clusters could be created to increase the performance of person-independent models without requiring a substantial amount of data. Furthermore, a decision level smoothing logic might be needed to improve the performances of these systems. If we can examine the labels assigned by machine learning algorithms from a high-level perspective, some errors could be noticed and corrected. For example, in one-minute frames, if the labels of consecutive frames are assigned as SRS (S: Stressed, R: Relaxed) respectively, we could determine that the middle relaxed label is a mistake because a person cannot relax and stressed in one minute. By applying decision level smoothing techniques, we can increase the classification accuracies of real-life stress detection systems.

After detecting high levels of enduring stress, these systems should also decrease the stress of individuals to the acceptable levels. When we examine the literature, there are very few number of studies dealing with managing stress in the daily life. Chen et al. [36] recognized mental stress by using heart activity collected from Zephyr wearable sensor. After they detected stress levels, they found breathing patterns for each user to help them relax. This is a YOGA respiratory pattern which has the most resemblance to individuals respiratory pattern. Akmandor et al. [26] designed a stress detection and reduction scheme. They collected ECG, EDA, blood pressure, respiration and blood oximeter data in the laboratory. For the reduction of the stress levels, classical music, warm stone, good news were applied. They showed that stress management schemes help people returning to the baseline state faster. Researchers developed a stress detection method and tested in the laboratory with nine participants [12]. They used both wrist-worn and chest-based heart activity sensors and finger-based electrodermal activity sensors. They divided the tasks into stressor (ice bucket, singing, scary game, SCWT (The Stroop Color and Word Test) and arithmetic) and non-stressor (conversation, eating email reading and homework) classes. They demonstrated that non-stressor tasks especially eating help participants relax. There are also ancient methods for alleviating stress. However, they generally require outdoor environments. An ideal stress management method should be applicable indoor, not require extra hardware or equipment and scientifically validated. Several mobile apps have closer properties to an ideal system such as Pause, HeartMath and Calm. However, the effect of these indoor applicable apps are not examined comprehensively but preliminary research has commenced [3]. There is also a need to apply indoor applicable relaxation methods and observe their effectiveness by examining physiological signals.

III. SYSTEM METHODOLOGY

In this section, we will investigate the performance of our stress detection scheme in two different manners. The first one is using the known context as the ground truth. We enumerated the different states as 1: baseline, 2: lecture (cognitive load), 3: exam (stress) and 4: recovery (stress management) states. We further provide these labels as classes to the machine learning algorithm. The second way is to use the perceived stress levels collected from self-reports as the ground truth. In order to measure the perceived stress levels, we collected NASA-TLX questionnaires from the participants. It includes mental demand, physical demand, temporal demand, performance, frustration and effort of participants from a session. We restructured NASA-TLX more appropriate for stress level detection studies by modifying the weights of the survey parts. The perceived level part is examined in the psychological approach whereas the physiological part is examined in the physiological approach subsections.

A. PHYSIOLOGICAL APPROACH

In this research, a stress detection system which uses heart activity, skin conductance, accelerometer and skin temperature for recognizing multiple stress levels. In order to eliminate the artifacts caused by unlimited movements in real-life scenarios, preprocessing tools specific for each modality were developed and used. The most distinctive features used in the literature for each signal were also selected and extracted. After the feature extraction phase, the best performing classifier algorithms were applied to the feature set. Our system works with the data collected from both Samsung Gear S smartwatches and Empatica E4 smart bands even though they have different software platforms and sensor types. As mentioned previously, for each modality we developed and used specific preprocessing and feature extraction tools. Each of them is described in detail.

1) EDA PREPROCESSING ARTIFACT DETECTION AND REMOVAL METHODS

Intense physical activity and temperature changes contaminate the EDA signal. Therefore, affected segments should be filtered out from the original signal. In order to detect the artifacts in the EDA signal, we used the EDA Explorer software [37] which is developed specifically for this signal. While developing this tool, experts labeled the artifacts manually. They trained an SVM model by using the labeled data. Accelerometer and skin temperature signals are also used for artifact detection. EDA Explorer achieved 95% accuracy for detecting artifacts. We removed the parts that this tool detects as artifacts from our signals. We further added batch processing and segmentation to this tool as new properties (see Figure 2).

2) EDA FEATURE EXTRACTION METHODS

After the artifact removal phase, features are extracted from the EDA signal. This signal has two components phasic and
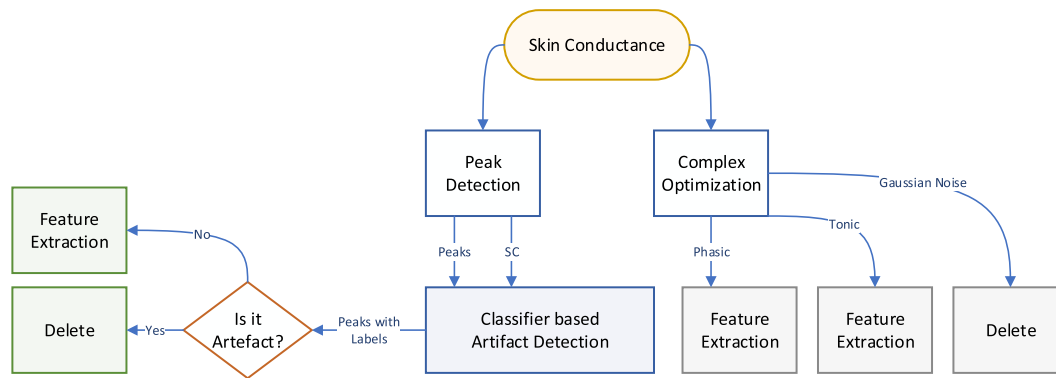
Y. S. Can *et al.*: Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory

IEEE *Access*

FIGURE 2. The detailed Skin Conductance processing module for artifact removal and feature extraction. EDAExplorer tool finds both peaks and artifacts separately. We selected the peaks that are not artifact by using this tool. After the removal of them, tonic and phasic components are decomposed and features from these components are extracted.

tonic namely. Features from both components are extracted. We used the cvxEDA tool [38] for decomposing the signal into these components. This tool uses convex optimization to estimate the Autonomic Nervous System (ANS) activity which is based on Bayesian statistics. After this tool is applied, our algorithm extracted features of both components. We selected the most discriminative features used in the literature [19], [39] and [40].

a: TONIC COMPONENT FEATURES

The tonic component in the EDA signal represents the long term slow changes. This component can be also called as the skin conductance level (SCL). It could be regarded as the indicator of general psychophysiological activation [41] and can depend highly on individuals [41]. The values generally can rise up to 15ms and above 20ms values are regarded as highly unlikely [41]. The tonic component is used since long term changes should not be overestimated with event-related fast changes. For this purpose, the phasic part is subtracted from the EDA signal. After the decomposition of EDA signal, we extracted mean, standard deviation, 20 percentile, 80 percentile and quartile deviation (75 percentile - 25 percentile) which are the most distinctive features in the literature [19], [39] from the tonic component (see Table 2).

b: PHASIC COMPONENT FEATURES

The phasic component represents faster (event-related) differences in the EDA signal. Peaks of EDA that happens as a reaction to a stimulus is also called Skin Conductance Response (SCR) [41]. It happens with a delay after the stimulus [42]. After we decompose the phasic component from the EDA signal peak related features e.g. peak per 100 seconds, a strong peak per 100 seconds are calculated. The peaks with more than 1 microSiemens response are identified as strong peaks. We chose this value by adapting the 1.5 microSiemens threshold in [19] to our data, since our peak range is lower than theirs.

3) HEART ACTIVITY PREPROCESSING ARTIFACT DETECTION AND REMOVAL METHODS

Unlimited movement of subjects and improperly worn devices also contaminates the heart activity signal collected from smartwatches and smart bands. In order to address this issue, we developed an artifact handling tool in MATLAB which has the batch processing capability. First, the data is divided into 50% overlapping segments [31] as recommended in the related literature. Artifact detection percentage rule (also employed in Kubios [43]) is applied after the segmentation phase. In this rule, each data point is compared with the local average around it. If the difference is more than a predetermined threshold percentage, the data point is labeled as an artifact. The threshold is defined as 20% difference which is commonly selected in the literature [31]. In our tool, we further developed two options after an artifact is detected: interpolation or further filtering. We described these methods in detail (see Figure 3).

a: ARTIFACT DETECTION PERCENTAGE THRESHOLD - REMOVAL

The first option is removing the artifact data point from the signal. However, after the data is removed, this creates holes in segments which makes it difficult to evaluate them as a whole. In order to overcome this issue, we implemented two filters: minimum consecutive time and minimum consecutive number of samples. The minimum consecutive time constraint dictates a minimum non-interrupted (with deleted artifact holes) time series with determined length on a segment to be evaluated and to extract features. Similarly, the minimum consecutive number of samples filter dictates a determined number of consecutive samples. By applying these filters, we ensure that segments that have too many artifacts distributed among the clear data are not evaluated and affect the performance of our system.

b: ARTIFACT DETECTION PERCENTAGE THRESHOLD - INTERPOLATION

After detecting the artifacts, another option would be to replace them with interpolation. The choice of the

TABLE 2. Heart rate variability, Electrodermal activity (EDA) and Acceleration features and their definitions.

| Feature	Description
Heart Rate Variability Features	
Mean RR	Mean value of the RR intervals
STD RR	standard deviation of the inter-beat interval
RMSSD	Root mean square of successive difference of the RR intervals
pNN50	Percentage of the number of successive RR intervals varying more than 50ms from the one
HRV triangular index	Total number of RR intervals divided by the height of the histogram of all RR intervals measured on a scale with bins of 1/128 s
TINN	Triangular interpolation of RR interval histogram
LF	Power in low-frequency band (0.04-0.15 Hz)
HF	Power in high-frequency band (0.15-0.4 Hz)
LF/HF	Ratio of LF-to-HF
pLF	Prevalent low-frequency oscillation of heart rate
pHF	Prevalent high-frequency oscillation of heart rate
VLF	Power in very low-frequency band (0.00-0.04 Hz)
SDSD	Related standard deviation of successive RR interval differences
Acceleration Features	
Mean X	Mean acceleration over x axis
Mean Y	Mean acceleration over y axis
Mean Z	Mean acceleration over z axis
Mean ACC MAG	Mean acceleration over acceleration magnitude axis
Energy	FFT energy over mean acceleration magnitude
Electrodermal Activity Features	
Mean Tonic	Mean of the phasic component
SD Tonic	Standard deviation of phasic component
Perc20	20th percentile of the phasic component
Perc80 Tonic	80th percentile of the phasic component
Quartdev Tonic	Quartile deviation (75 percentile - 25 percentile) of the phasic component
Strong Peaks Phasic	The number of strong peak per 100 seconds
Peaks Phasic	The number of peaks per 100 sec.
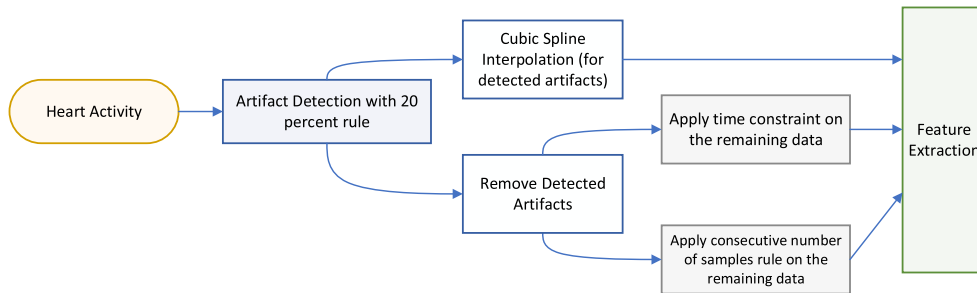
FIGURE 3. The detailed Heart Activity processing module for artifact removal and feature extraction. First, the artifacts are detected and removed. After the removal, the algorithm provides two different options. The first one is the interpolation of the removed points. The second option is to apply minimum consecutive time and sample constraints on the remaining data to be regarded as meaningful. After preprocessing, HR features are extracted.

interpolation technique is a critical decision. The interpolation function should be similar to the heart signal. To this end, we selected shape preserving cubic spline interpolation and applied the built-in MATLAB function. The tool has further batch processing feature. Parameters such as length of local mean, the percentage of artifact detection rule, minimum consecutive time and data sample constraints are parameters that could be changed in our tool.

4) HEART ACTIVITY FEATURE EXTRACTION METHODS

In order to extract features from the heart activity signal, MATLAB built-in tools and Marcus Vollmer's HRV toolbox [44] are used. The features could be examined in time and frequency domain categories. These particular time and frequency domain features are chosen because they are widely used as the most discriminative ones in [19], [39] and [40].

Y. S. Can *et al.*: Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory

IEEE *Access*

a: TIME DOMAIN FEATURES

We searched the literature and selected the most distinctive features in the time domain. Mean value of the heart rate (Mean HR), standard deviation of the inter-beat interval (STD RR), mean value of the inter-beat (RR) intervals (Mean RR), root mean square of successive difference of the RR intervals (RMSSD), the percentage of the number of successive RR intervals varying more than 50ms from the previous interval (pNN50), the total number of RR intervals divided by the height of the histogram of all RR intervals measured on a scale with bins of 1/128 s (HRV triangular index), and triangular interpolation of RR interval histogram (TINN) are selected and extracted from the heart activity signal.

b: FREQUENCY DOMAIN FEATURES

Features from the frequency domain are also extracted. However, since the heart peaks are not equidistant from each other, FFT could not be applied directly. We first preprocess the signal for equal distant samples (resample) and then applied FFT. We further applied Lomb-Scargle periodogram [45] which is developed for this type of signal to convert to the frequency domain. We extracted features from both methods. Low frequency power (LF), high frequency power (HF), very low frequency power (VLF), prevalent low frequency (pLF), prevalent high frequency (pHF), the ratio of LF to HF (LF/HF) (Preprocessing + FFT), LF, HF, LF/HF (Lomb-Scargle) features are selected and extracted from frequency domain representation of the heart activity signal.

5) ACCELEROMETER FEATURE EXTRACTION METHODS

The accelerometer sensor data is used for two different purposes. Firstly, we extracted features from this sensor. As mentioned above, this sensor was also employed to clean the EDA data along with the skin temperature sensor as a second use. The mean value of 4-axis and the frequency domain energy of magnitude were the extracted features.

6) DATA FUSION

After we divided our data into segments (between 60-480 seconds) different modalities should be merged. We chose these segment sizes because the duration of stress stimulation and recovery processes is around a few minutes [46]. Therefore, by using these values we expect to capture a stress event in a segment. Especially, heart activity signal starts with a delay (to calculate the heartbeat per minute at the start) and all signals should be synchronized. We included start and end timestamps for each segment and each modality is merged with a script if their time intervals are overlapping.

7) PREPROCESSING OF DATA AND HANDLING OF IMBALANCE BEFORE CLASSIFICATION

Our data set is not balanced when the number of instances belongs to each class is considered. About 40% of frames belong to cognitive load, 20% of frames belong to stress and 20% of frames belong to recovery and 20 % of frames belong to baseline classes. We balanced the set by removing extra samples from the majority classes. Therefore, we prevented classifiers from biasing towards the class with more instances. We have a total of 1800 frames which gave us the option to remove extra frames of the cognitive load class. We also examined the effect of Synthetic Minority Over-sampling Technique (SMOTE) [47] in Table 14. The Weka toolkit [48] has several preprocessing features before classification. Numeric to nominal transformation among them was used to convert the stress level column into nominal class attribute. We further normalized the features to prevent overfitting.

8) MACHINE LEARNING CLASSIFIER ALGORITHMS

The classification algorithms in Weka toolkit [48] are used for discrimination of stress from the cognitive load. It is the most commonly used and comprehensive machine learning platform in the literature.

In this study, we employed five different machine learning classification algorithms to recognize different stress levels:
a) Principal component analysis (PCA) and support vector machine (SVM) with a linear kernel
b) Random Forest (with 100 trees)
c) K-nearest neighbours (n = 1)
d) PCA and Linear discriminant analysis (LDA)
e) MultiLayer Perceptron (A Shallow Neural Network) (MLP)

These classifiers are chosen because they are the most prominent traditional ML algorithms. MLP (Shallow Neural Network), Random Forest, kNN (One of the Simplest), SVM and LDA are the most commonly used and successful classification algorithms for stress detection [19], [31]. The data is divided into 90% training and 10% test parts. 10-fold cross-validation was applied. Our system is capable of detecting stressed, cognitively loaded and baseline states by using the mentioned algorithms.

B. PSYCHOLOGICAL APPROACH

1) NASA-TLX

The NASA Task Load Index (NASA-TLX) is an assessment tool that aims to evaluate the perceived workload. Usually, this tool is utilized for the assessment of performance tasks in specific operational environments [49]. Originally developed for use in aviation, it has been later implemented in different contexts, particularly in human factors research. It is a six-dimensional scale from one or more subject while they are performing a task. These scales are mental demand, physical demand, temporal demand, performance, effort and frustration. The NASA-TLX has two parts. In the first part, the subject is directed to give a score between 0 and 100 to each of these scales. This part has six questions. In the second part, the questionnaire is intended to create a weight for each scale. The subject is directed to make a pairwise comparison for each scale. The second part has 15 questions. Our subjects from a previous study gave negative feedback about the
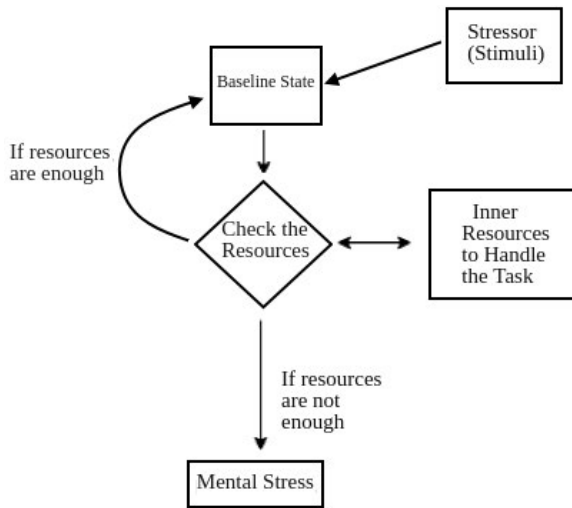
IEEE *Access*

Y. S. Can *et al.*: Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory

FIGURE 4. When the stressor appears, people check their inner resources. The stressor could be mental demand, physical demand, temporal demand, etc. If they feel that they can handle these demands, they will not feel stressed. Otherwise, they will have mental stress.

length of the questionnaire since it has 21 questions in total. Therefore, we used ''RAW-TLX'' by not applying the second part of the NASA-TLX questionnaire. Each scale is examined throughout the study.

2) ADAPTING NASA-TLX FOR STRESS DETECTION STUDIES

The NASA Task Load Index (TLX) [50] is comprised of six subscales mentioned above. Taken together, these subscales are aimed to operationalize the construct of mental workload, and thus the NASA-TLX has principally been implemented in organizational settings ([51] and [52] just to mention two examples). It is designed to measure the estimated workload of subjects while they are performing a task or immediately afterward [53].

Some studies used the raw NASA-TLX for measuring stress levels [22], [25] and [15]. However, it is clear that the workload is not the same as stress. In other words, cognitive load in general and cognitive workload in particular do not represent stress (represent perceived stress even less) as a construct, which is mainly defined as the subjective appraisal of the situational context, including the weighing of the own resources to cope with a certain stressor. The first 5 subscales of the NASA-TLX are only partially overlapped with the conceptualization of what comprises a stressor. Especially the physical demand component of the raw NASA-TLX becomes irrelevant in mental stress detection. Furthermore, high workload does not necessarily mean high stress levels. If the subject is confident to deal with the workload with his/her own resources, it might not cause stress (see Figure 4). Too many difficult mathematics questions do not create stress on experts in the field. On the contrary, the sixth subscale, frustration, can partially represent subjective stress, understood as the extent to which a person finds a situation to exceed his capacity to successfully deal with the situation.

It includes information about insecurity, discouragement, irritation, stress of individuals. For this reason, and given that we wanted to have a perceived stress measure, we have only taken into consideration the last subscale.

IV. EXPERIMENT DESIGN
A. ILKYAR SUMMER SCHOOL SEMINARS FOR TEACHERS
Every year, teachers from public schools of different cities in Turkey are gathered and participate in seminars which are given by university lecturers. This year, thirty-two teachers participated in the ILKYAR summer school. A seminar session took approximately three hours. We collected data during this session with wrist-worn wearable devices which are the combination of Samsung Gear S2, Samsung Gear S and Empatica E4.

We first described the general outline of the study and delivered informed consent forms to the teachers. The experiment started after volunteering participants signed these forms. All participants wore the wrist-worn wearables and turn on the devices for data collection. At the beginning of the event, Pittsburgh Sleep Quality Index (PSQI), General Wellbeing Index (GWBI), WHO-five Well-being Index (WHO-5) Well Being and Perceived Stress Scale (PSS) baseline questionnaires were collected. We explained these questionnaires to the participants and baseline signals are recorded afterwards.

In the baseline session, participants read neutral magazines about sports, design, clothes and cars. It took about 20 minutes. Following the baseline session, the lecture about the research in the Computer Engineering Department was briefly given. The length of the lecture session was about 40 minutes. We gave a break after the lecture which was about 10 minutes. When they returned, we told them there is going to be an important exam and we measure their performance. The exam consisted of arithmetic tasks inspired by Trier Mental Stress Test. The exam lasted for 20 minutes. The last part of the experiment was the recovery session. They listened to some relaxation music, they are told to take deep breathes and think of their positive memories. In this way, we tried to reduce their stress levels with the method inspired by the HeartMath app [54]. The recovery sessions also took approximately 20 minutes. After all sessions (baseline, lecture, exam and recovery), raw NASA-TLX questionnaires are collected from the participants about the sessions. The procedure of the experiment is shown with the chronological order in Figure 5:

B. PARTICIPANTS AND APPARATUS
All participants are public school teachers within the age range of 25 to 40. There were 10 female and 22 male participants. For the wearable devices used in the experiment, we searched the market with certain criteria. The devices should be non-obtrusive, provide the raw physiological data (with official SDK), give the Inter-Beat-Interval (IBI) from heart activity. IBI is used for heart rate variability (HRV) measurement. HRV is an important indicator
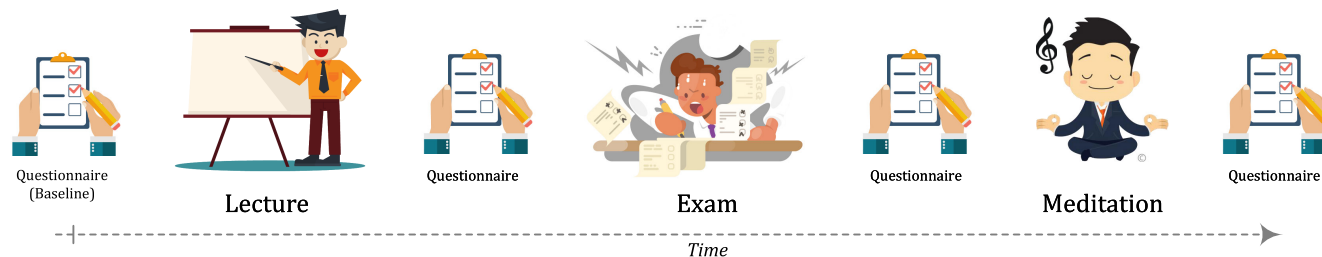
Y. S. Can *et al.*: Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory

IEEE *Access*

FIGURE 5. The timeline of the event is demonstrated. After each session, self reports were collected. We further took the baseline questionnaires at the beginning of the event.

of stress. An optional feature would be providing EDA data. EDA is another important physiological signal for especially arousal detection. Empatica E4 satisfied all our expectations. We also investigated the Samsung branded Gear S1, S2, and S3 smartwatches. However, with the S3, Samsung stopped providing IBI intervals with its original SDK. Previous Samsung smartwatches (S1, and S2) give access to IBI raw data. Microsoft Band 2 has also the two signals. However, raw data access is not provided with the original SDK. It is removed from the official website. Apple smartwatches also do not provide raw data. We further evaluated the Polar chest band. However, it could be considered as an obtrusive device for real-life settings. Therefore, Empatica E4, Samsung S1 and Samsung S2 off-the-shelf devices were selected for the experiment.

When the battery lives of the devices were compared, Empatica E4 outperformed Samsung S smartwatches. It collects data for approximately two days. On the other hand, Samsung smartwatches can collect data for approximately 4 hours when all sensors are active. However, it is important to note that Empatica E4 is developed for research purposes and it is more expensive than Samsung commercial smartwatches. Samsung devices provide data via the Bluetooth connection. Conversely, Empatica E4 have cloud support and data could be downloaded from the cloud server. Data acquired by both devices are available in CSV format.

As far as sensors are concerned, Empatica E4 devices have four sensors for measuring the acceleration, photoplethysmography, electrodermal activity and the skin temperature while Samsung Gear watches lack EDA and skin temperature sensors but instead they are equipped with Gyro and Barometer sensors. In this study, we used PPG, EDA and ACC sensors for feature extraction.

C. ETHICS

The procedure of the methodology used in this study is approved by the Institutional Review Board for Research with Human Subjects of Bogazici University with the approval number 2018/16. Each subject signed a consent form explains the procedure of the experiment and its aims and implications to both the society and the subject before the data recording begins. We further described the procedure verbally to the participants. The recorded data are stored anonymously.

D. DATA DESCRIPTION

Data is recorded in different csv files in both devices with timestamps. Participants are differentiated with the used device id. Data is stored in different folders for each device. For each modality, a different file is created under these folders. A sample figure for different modalities of our data is shown in Figure 6.

1) EDA DATA
EDA data is collected with Empatica E4 devices. The format of this file is starting with the Linux timestamp followed by samples. Empatica stated the sampling frequency of EDA as 4Hz. EDA signal is used for feature extraction purposes.

2) IBI DATA
IBI data can be collected with all the wearable devices we used in this experiment. All devices provide two columns for this type of data. The first column includes the IBI interval and the second column points out the time of the sampling. This type of data structure is necessary because non-equidistant sampling is employed due to the changing frequency of heartbeats in time. We used this physiological data type for feature extraction after the artifact removal and interpolation.

3) ACC DATA
ACC data can be also recorded with all type of mentioned devices in this experiment. There are three columns which indicate X, Y and Z coordinates of the acceleration. The sampling rate is 64Hz. We used this physiological data type for both feature extraction and EDA artifact removal.

4) ST DATA
Skin Temperature (ST) data is used for EDA artifact detection and removal. Only Empatica E4 devices have the capability to collect this modality and the sampling frequency is 4Hz.

V. DISCUSSION
A. PSYCHOLOGICAL APPROACH
In this section, we examined two different issues. We conducted an experiment that consists of lecture, exam and recovery sessions. We first investigated the success of our experiment in creating three different psychological states by
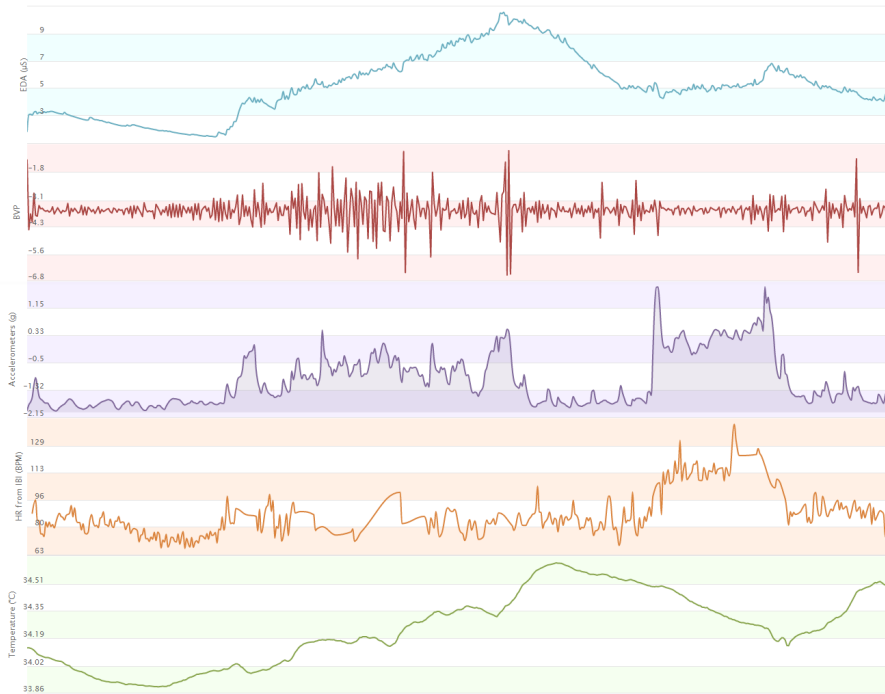
IEEE *Access*

Y. S. Can *et al.*: Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory

FIGURE 6. A sample figure for the collected data with an Empatica E4 device. Different modalities are shown separately.

using the collected raw NASA-TLX self-reports. T-test was applied to measure the separability of these sessions from each other. We further applied the same procedure after modifying raw NASA-TLX to demonstrate the separability when the modified version is used. By doing that, we showed that our experiment creates three different psychological states and raw NASA-TLX and modified version can show the difference between these states.

1) CLUSTERING OF WORKLOAD AND CONTEXT
In this study, we examined the workload in different contexts for our experiment. As mentioned above, NASA-TLX is a tool which can be used to evaluate the perceived workload. By examining self-reports collected from participants after lecture, exam and recovery sessions, the success of inducing different workloads for these sessions is investigated. The t-test in the R programming language is used to measure whether these sessions are different or not in terms of perceived workload. The paired t-test is used to evaluate the separability of each session. The degree of freedom is 31. We applied the variance test to each session tuple, we could not identify equal variance in any of the session tuples. Thus, we selected the variance as unequal. We used 95% confidence intervals. The t-test results are provided in Table 3. For all tuples, the null hypotheses stating that recovery is greater than or equal to lecture, lecture is greater than or equal to exam and recovery is greater than or equal to exam sessions are rejected. The following p-values and test statistics are provided in Table 3. The perceived workload levels of

TABLE 3. T-test results for session tuple comparisons of perceived workload using RAW-TLX.

Session Tuple	Test statistic	p-value
Recovery - Exam	-13.073	$1.869e^{-14}$
Recovery - Lecture	-3.5886	0.0005645
Lecture - Exam	-9.723	$3.121e^{-11}$

participants for the exam, lecture and recovery sessions are observed to be significantly different.

From low to high perceived workloads, it can be sorted as the following: *Recovery < Lecture < Exam*. The boxplot for session tuples and all sessions are provided in Figure 7.

2) CLUSTERING OF SURVEY DATA AND CONTEXT AFTER NASA-TLX MODIFICATION
After we modified the NASA-TLX for measuring perceived stress levels, we further examined the self-reports in different sessions. In other words, if our stress induction, cognitive load induction and recovery sessions are successful in terms of creating different perceived stress levels, self-reports for each session should be separable. The same methods with Section 5.1.1 are applied for perceived stress levels. The results are provided in Table 4. The perceived stress levels of the t-test are sorted from low to high as *Recovery < Lecture < Exam*. Recovery session is designed as lower/mild stress and lecture also creates a cognitive load which is close to mild / low stress tasks in terms of perceived stress. The similarity between these sessions is demonstrated with

Y. S. Can *et al.*: Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory

IEEE *Access*

TABLE 4. Paired T-test results for session tuple comparisons of perceived stress using frustration score.

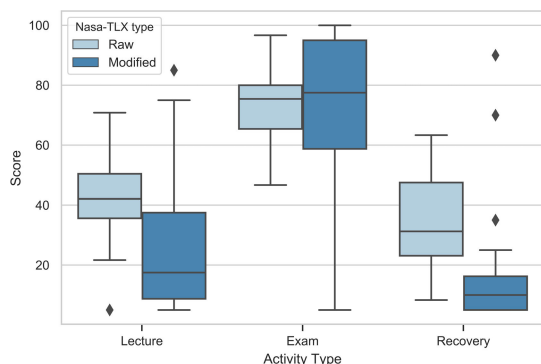| Session Tuple	Test statistic	p-value
Recovery - Exam	-8.3929	$1.762e^{-9}$
Recovery - Lecture	-2.6391	0.01289
Lecture - Exam	-6.7003	$1.702e^{-7}$

FIGURE 7. The box plots of the raw and modified Nasa Task Load Index (NASA-TLX) self-report scores during recovery, examination and lecture sessions.

the t-test. The boxplot for session tuples and all sessions are provided in Figure 7.

For all tuples, the null hypotheses states the above comparisons of Section 5.1.1 are rejected for perceived stress. The following p-values and test statistics are provided in Table 4. The perceived stress levels of participants for the exam, lecture and recovery sessions are also determined to be significantly different and from low to high perceived stress, it can be sorted as the following, *Recovery < Exam < Lecture*. The boxplot for session tuples and all sessions are provided in Figure 7. As in the case of perceived workload, all of the mean value of the distributions are different, however in terms of distribution recovery and lecture sessions are the most similar tuples. Recovery session is designed as lower/mild stress and lecture also creates a cognitive load which is close to mild / low stress tasks in terms of perceived stress. The similarity between these sessions is expected.

3) CLUSTERING OF PARTICIPANTS WITH THEIR BASELINE SURVEYS

Our 32 participants filled the Perceived Stress Scale – 14 (PSS14) questionnaires at the beginning of the event. We calculated the scores of each individual. This questionnaire consists of questions regarding the stress level felt during the last month before the experiment. By using these scores, we clustered the participants into low stress, medium stress and high stress classes. These classes demonstrate the condition of participants before the experiment. The top 33 percentile is assigned to high stress, low 33 percentile is assigned to low stress and the remaining 34 percentile is assigned to medium stress classes. We used these clusters for developing specific models in the next section.

TABLE 5. Effect of number of different modalities and combination of them on the system performance. Note that number of classes are fixed at 2 (stressed and recovery) and window size is 60 seconds.

Algorithm	Accuracy		
	HR	**EDA**	**Combined**
MLP	76.37	62.78	79.88
RF	82.70	81.82	85.63
kNN	73.87	73.29	74.13
LDA	74.15	61.36	68.39
SVM	79.54	66.19	68.96

B. PHYSIOLOGICAL APPROACH

1) EFFECT OF DIFFERENT MODALITIES TO STRESS LEVEL CLASSIFICATION ACCURACIES USING KNOWN CONTEXT

In this study, we investigated the effect of two different modalities on classification accuracies. We further examined the effect of combining these two modalities. Heart rate variability (HRV) and electrodermal activity (EDA) are among the most discriminative physiological signals for stress level detection. HRV achieved higher classification accuracies than EDA with all the classifiers. From this result, we can deduce that HRV is a more discriminative physiological signal than EDA. Another important finding was the combination of these two signals achieve higher stress level classification accuracies than these signals alone in most cases. Combination of modalities has a positive effect on the stress level detection accuracy (see Table 5).

2) PERCEIVED STRESS, WORKLOAD LEVEL DETECTION FROM SELF-REPORTS AND PHYSIOLOGICAL STRESS LEVEL CLASSIFICATION IN KNOWN CONTEXT

We measured physiological stress and perceived stress separately in this study. Normally, it is expected that when individuals exhibit signs of high physiological stress, they also have high perceived stress levels. However, this might not be the case in some situations [1]. Therefore, the two stress types are examined independently. The ground truth in physiological stress was selected as the stress level of known context. The perceived workload levels are measured with the answers from the whole NASA-TLX questionnaire. In perceived stress, we used the stress level obtained from the modified NASA-TLX self-report answers. The answer scale collected in lecture, exam and recovery sessions is divided into three levels. The low stress level is between 0 and 30 in NASA-TLX self-report (scaled from 0-100), the medium stress level is assigned if the answer is between 35-75 and the high stress level is assigned if the self-report is equal or above 80. We first examine the performance of modified the NASA-TLX questionnaire for measuring perceived stress levels. The correlation with known context stress labels increased from 0.32 to 0.54 with the modified NASA-TLX when compared with whole NASA-TLX. This justifies the modification for better perceived stress level measurement. Since we collected answers in lecture, exam and recovery sessions, we observed that low and mild stress levels are seen in lecture and recovery sessions, thus we combined recovery

IEEE Access

Y. S. Can *et al.*: Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory

TABLE 6. Physiological stress, perceived workload and stress detection accuracies, EDA signal, the number of distinguished classes is 2 (recovery - cognitive load (low / mild stress), exam (high stress)) and windows size is 240 seconds. The NASA-TLX Perceived Workload and Perceived Stress scores are divided into three classes. Two classes (low and mild) in lecture and recovery sessions are combined in both known context and perceived workload and stress evaluation.

Algorithm	Accuracy with different Ground Truth Labels		
	Physiological Stress	Perceived Workload	Perceived Stress
MLP	77.00	92.00	88.50
RF	82.00	94.50	90.00
kNN	78.50	93.00	89.50
LDA	78.00	92.00	89.50
SVM	77.00	91.00	88.00

TABLE 7. Physiological stress, perceived workload and stress detection accuracies, HR signal, the number of distinguished classes is 2 (recovery - cognitive load (low / mild stress), exam (high stress)) and windows size is 240 seconds. The NASA-TLX Perceived Workload and Perceived Stress scores are divided into three classes. Two classes (low and mild) in lecture and recovery sessions are combined in both known context and perceived workload and stress evaluation.

Algorithm	Accuracy with different Ground Truth Labels		
	Physiological Stress	Perceived Workload	Perceived Stress
MLP	78.24	92.64	86.23
RF	84.19	94.52	89.51
kNN	79.96	92.33	85.29
LDA	78.56	94.21	88.42
SVM	78.87	94.52	88.73

and lecture into the low/mild stress class. In the analysis of self-reports in Section 5.1.2, the two classes have closer behavior as expected. The exam session is determined as the high stress class. The same procedure is also applied to the known context ground truth labels and recovery and lecture sessions are merged into one low / mild stress context. The exam session is labeled as the high stress context. The perceived workload, perceived stress and physiological stress detection accuracy results are provided in Tables 6 and 7. The perceived workload results have always the best performances with all classifiers. The workload level (physical, mental, temporal demand, performance, effort and frustration) could be easily differentiated by individuals with self-reports. The prediction accuracy of the perceived stress is weaker than the perceived workload. Individuals may not always perceive the true stress levels or express them with self-reports [1] and these issues might result in the appearance of lower system performance. The lowest classification results are obtained when measuring physiological stress levels with known context ground truth labels. Although we prove the success of our stressors from the self-reports, some participants could not be induced with the desired levels of stress or cognitive load. Our data collection did not take place in a laboratory. Instead, we record data in a real-life event. Thus, some participants may not be cognitively loaded in lectures or stressed in exams in a semi-controlled real-life event and this decreases the accuracy of stress detection accuracy with known context ground truth labels.

TABLE 8. Effect of general, personalized and clustered models on system performance, EDA signal. Note that number of distinguished classes is 3 (relax, cognitively loaded, stressed) and window size is 120 seconds.

Algorithm	Models		
	Personalized	Cluster	General
MLP	74.25	62.55	59.50
RF	74.45	70.57	64.46
kNN	71.50	66.69	57.44
LDA	70.18	58.91	58.88
SVM	75.04	63.64	63.64

TABLE 9. Effect of general, personalized and clustered models on system performance, HR signal. Note that number of distinguished classes is 3 (relax, cognitively loaded, stressed) and window size is 120 seconds.

Algorithm	Models		
	Personalized	Cluster	General
MLP	79.87	70.08	66.69
RF	81.16	74.87	71.57
kNN	76.17	67.73	59.37
LDA	78.08	67.96	66.80
SVM	74.94	70.82	66.26

3) EFFECT OF DIFFERENT MODELS: PERSONALIZED, CLUSTER-SPECIFIC AND PERSON-INDEPENDENT MODELS

Since the stress reaction of individuals has a unique pattern, the ideal way to develop automatic stress detection models is to use the individual's data. However, in most cases there is not enough personal data for developing this kind of model. Another way is to develop models from all collected data and apply this one model to all people. However, the accuracy of this model is expected to be lower than the personalized model because of the mentioned person-specific stress reactions. In this section, we offered a hybrid approach. As mentioned in Section 6.1.3, we clustered participants by using their baseline stress levels. By using self-report answers regarding the month before the experiment, we divided them into low stress, medium stress and high stress clusters. We then develop models for each cluster separately since we expect people in the same cluster might have similar physiological reactions to our stimuli. As it can be seen in Tables 8 and 9, person-independent models have the lowest stress classification accuracies whereas personalized models obtained the best results with all classifiers. Our hybrid models have accuracies lower than personal and higher than person-independent models. When the data is not enough for personal models, our hybrid approach could be used to increase the performance of the system.

4) EFFECT OF STRESS DETECTION INTERVAL AND RESOLUTION TO CLASSIFICATION ACCURACIES

Another important research issue that we want to address is to find the optimal interval for stress detection studies. In other words, since stress reaction has certain physiological characteristics, there might be an optimum time interval that the stress level could be detected more easily from biofeedbacks of individuals. We carried out experiments with 60,

Y. S. Can *et al.*: Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory

IEEE *Access*

TABLE 10. Effect of stress resolution to stress detection accuracies. Number of distinguished classes is fixed at 2(recovery, stressed), EDA signal.

Algorithm	Accuracy in different stress detection resolutions			
	60s	120s	240s	480s
MLP	62.78	69.04	73.61	69.56
RF	81.82	75.00	75.00	52.17
kNN	73.29	73.80	72.22	56.52
LDA	61.36	59.52	66.67	65.21
SVM	66.19	69.04	76.38	73.91

TABLE 11. Effect of stress resolution to stress detection accuracies. Number of distinguished classes is fixed at 2(recovery, stressed), HR signal.

Algorithm	Accuracy in different stress detection resolutions			
	60s	120s	240s	480s
MLP	76.37	76.56	75.22	60.24
RF	82.7	83.79	78.76	73.49
kNN	73.87	76.95	77.43	61.44
LDA	79.54	79.68	81.42	71.08
SVM	74.15	76.75	72.56	65.06

TABLE 12. Effect of number of stress levels to stress detection accuracies. Note that window size is fixed to 120 seconds and enumerated classes are as follows: 1 (cognitive load - lecture), 2 (relax), 3 (stressed- exam), 4 (recovery- stress management), EDA signal.

Algorithm	Classes				
	1 vs 2	1 vs 3	3 vs 4	2 vs 3	1 vs 2 vs 3
MLP 1-HL	73.94	49.52	69.04	77.63	59.50
MLP 2-HL	66.72	46.27	59.52	75.42	51.20
RF	78.16	72.58	75.00	83.42	64.46
kNN	73.16	64.42	73.81	78.16	57.44
LDA	71.31	50.96	61.31	76.05	58.88
SVM	73.42	55.77	59.52	76.84	59.92

TABLE 13. Effect of number of stress levels to stress detection accuracies. Note that window size is fixed to 120 seconds and enumerated classes are as follows: 1 (cognitive load - lecture), 2 (relax), 3 (stressed- exam), 4 (recovery- stress management), HR signal.

Algorithm	Classes				
	1 vs 2	1 vs 3	3 vs 4	2 vs 3	1 vs 2 vs 3
MLP 1-HL	82.25	65.83	76.17	74.43	66.69
MLP 2-HL	71.31	46.27	59.52	75.42	54.42
RF	83.53	73.35	84.76	80.33	71.57
kNN	69.71	68.03	76.95	73.37	59.37
LDA	68.72	65.83	76.75	75.79	66.26
SVM	83.15	69.43	79.68	77.91	66.80

TABLE 14. Effect of imbalance handling methodology: under-sampling of the majority class and SMOTE. Accuracy is reported on HR signal for discrimination of stressed-exam and relax states.

Classifier	Class Balancing Method	
	Under-sampling	SMOTE
MLP	74.43	60.10
RF	75.42	78.15
kNN	80.33	66.97
LDA	75.79	55.09
SVM	77.91	69.76

TABLE 15. Effect of number of stress levels to stress detection in terms of Area under Curve (AUC). Note that the window size is fixed to 120 seconds and enumerated classes are as follows: 1 (cognitive load - lecture), 2 (relax), 3 (stressed- exam), 4 (recovery- stress management), using the EDA signal.

Classifier	Classes				
	1 vs 2	1 vs 3	3 vs 4	2 vs 3	1 vs 2 vs 3
MLP w. 1-HL	0.643	0.458	0.604	0.625	0.531
MLP w. 2-HL (DNN)	0.534	0.421	0.523	0.580	0.531
RF	0.786	0.800	0.820	0.848	0.783
kNN	0.707	0.636	0.727	0.722	0.632
LDA	0.563	0.393	0.573	0.502	0.516
SVM	0.639	0.543	0.594	0.649	0.555

TABLE 16. Effect of number of stress levels to stress detection in terms of Area under Curve (AUC). Note that the window size is fixed to 120 seconds and enumerated classes are as follows: 1 (cognitive load - lecture), 2 (relax), 3 (stressed- exam), 4 (recovery- stress management), using the HR signal.

Classifier	Classes				
	1 vs 2	1 vs 3	3 vs 4	2 vs 3	1 vs 2 vs 3
MLP w. 1-HL	0.853	0.828	0.889	0.836	0.815
MLP w. 2-HL (DNN)	0.800	0.857	0.840	0.836	0.807
RF	0.896	0.805	0.934	0.865	0.886
kNN	0.789	0.758	0.853	0.780	0.779
LDA	0.868	0.811	0.847	0.690	0.770
SVM	0.800	0.687	0.726	0.724	0.721

120, 240 and 480 second intervals. In 9 out of 10 experiments (5 classifiers with EDA and 5 classifiers with HR), the best accuracies are found with 120 - 240 second intervals (see Tables 10 and 11). However, researchers should also take into account the employed classifier algorithm when determining the optimal interval for stress detection.

5) EFFECT OF NUMBER OF RECOGNIZED STRESS LEVELS TO CLASSIFICATION ACCURACIES

The effect of the number of recognized stress level classes to the accuracies is also examined. As mentioned, our experiment has four different sessions: baseline, lecture, exam and recovery with guided mindfulness. It is assumed that the lecture will induce a cognitive load, the exam will induce stress on the participants. We tried to bring them back to their baseline states by applying guided mindfulness with a relaxing music. We experimented with different tuples from these four sessions. Lastly, we examined the performance of our system on three class classification.

Three classes are selected as stressed, baseline and cognitive load. Lecture vs. stress is the most difficult to distinguish session tuple with both types of signals as it can be seen

in Tables 12 and 13. This is because of the similarity of physiological reactions of cognitive load and stress behaviors. Exam vs. recovery, lecture vs. baseline and exam vs. baseline session tuples could be differentiated with relatively higher accuracies with both modalities. Another important finding is that exam and recovery sessions can be distinguished with accuracies similar to (and sometimes higher than) lecture vs. baseline and exam vs. baseline session tuples. This shows that our recovery session successfully alleviates the stress of

IEEE *Access*

Y. S. Can *et al.*: Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory

TABLE 17. High level accuracy calculation and decision level smoothing accuracy results with EDA signal. Note that number of classes is fixed at 2 (stressed and recovery) and window size is 60 seconds.

Algorithm	Different decision level smoothing and high level Accuracy calculation		
	EDA	Decision Level Smoothing	HighLevelAcC
MLP	62.78	70.17	81.25
RF	81.82	92.89	100.00
kNN	73.29	86.07	93.75
LDA	61.36	62.78	68.75
SVM	66.19	70.17	81.25

TABLE 18. High level accuracy calculation and decision level smoothing accuracy results with HR signal. Note that number of classes is fixed at 2 (stressed and recovery) and window size is 60 seconds.

Algorithm	Different decision level smoothing and high level Accuracy calculation		
	HR	Decision Level Smoothing	HighLevelAcC
MLP	76.36	88.18	94.44
RF	82.70	92.12	90.74
kNN	73.87	87.32	90.74
LDA	74.15	85.30	85.30
SVM	79.54	89.62	87.03

participants and decrease their stress level. The three class classification accuracy is similar to lecture - exam tuple but less than other tuples. The difficulty in distinguishing these two sessions is also interfering with the performance of the three class classification system. However, even with these three classes, we have similar accuracies with reported systems differentiating stress from cognitive load in laboratory settings [55], [56].

We also tested the effect of class imbalance problem handling techniques: SMOTE and removal of the majority classes in Table 14. In most of the cases, SMOTE has lower accuracies than the other technique. It increases the accuracy of the RF classifier. In Table 15 and Table 16, we also provided the AUC results to compare the performance of classification systems. We tested MLP with different numbers of hidden layers. Deep Neural Networks (DNN) increase the performance of systems when there are a huge amount of data. This result showed that our data size might not be sufficient for DNNs to learn and create better models than the traditional classifiers.

6) INCREASING ACCURACIES WITH DECISION LEVEL SMOOTHING

The classification errors can be corrected by examining the results from a high level perspective. We examined our decisions for a 60 second interval with this perspective since the state of participants does not likely to change in such a short interval and it takes at least a few minutes for stimulation and recovery processes to complete [46]. We search for cases which are not likely to occur when logically evaluated and add some rules on top of our system. Our rule was correcting changes with unusually high frequency. In other words, if a subject is found out to be stressed in one window, not stressed in the consecutive one and stressed again in the next window; we determined this case as highly unlikely and an error

of our system. We applied our logic on top of our system automatically. The maximum accuracy of our system increase from approximately 82% to 92% with EDA and HR signals (see Tables 17 and 18). The performance of all classifiers increases significantly with decision level smoothing.

7) HIGH LEVEL ACCURACY CALCULATION FOR STRESS DETECTION

We divide all experiment data into 60 second windows and test each window separately when calculating the accuracy. However, in real-life, detecting stress for particular sessions and time intervals might gain more importance. Thus, we propose a different stress level detection accuracy calculation. For all sessions, we labeled all small windows and applied majority voting afterward for N consecutive intervals in a sliding window fashion. To put it another way, our system labels sessions by the majority of labels of small consecutive windows. We called this method as 'high-level accuracy calculation'. In this way, the accuracy for 2-class stress level detection goes up to 94.44% with HR signal and 100% with EDA signal (see Tables 17 and 18). If the aim is to identify stress levels in specific sessions, high-level accuracy calculation could be used to increase the performance.

VI. CONCLUSION AND FUTURE WORK

We proposed new models and methods for improving multi-level real-life stress detection systems using unobtrusive off-the-shelf smartwatches and smart bands. We tested our algorithms in real-life settings which include baseline, cognitive load, stress and recovery sessions of 32 participants in a summer school. First, the effect of our hybrid personal stress level clustering was examined. In the person-independent model, the data of all participants is divided into training and test parts. The personalized model uses the data of each participant for developing a model.

Y. S. Can et al.: Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory

IEEE Access

On the other hand, in our new hybrid model, we first cluster the participants into low, medium and high stress levels by examining their baseline self-reports. In this method, we develop specific models for each cluster. The personalized model has the highest and the person-independent model has the lowest accuracy. Our hybrid model has accuracies in between. It could be used in cases where there is not enough data of participants to develop personalized models. In these situations, our hybrid models will increase the accuracy of the system when compared with person-independent models.

Furthermore, the perceived stress, workload and physiological stress were investigated. We started with successfully classifying perceived workload level (3-class) using NASA-TLX. The minimum classification accuracy is 91% and the maximum accuracy is 94.52% for 3-classes. After that, we used the modified version of NASA-TLX to measure the perceived stress levels and compared with physiological stress levels. Modifying the NASA-TLX increased the correlation with known context labels from 0.32 to 0.54. When the performance of 3-level physiological and perceived stress detection classification accuracies are compared, perceived stress levels are always detected more successfully with all classifiers. Some participants might feel a different stress level than known context labels. This might decrease the performance of the physiological stress level detection system.

We further tested a decision-level smoothing method using the fact that stress levels of participants do not oscillate instantaneously. Our maximum accuracies with using a single modality are around 80% in 2-class classification (81.82% with EDA, 82.70%). To increase the performance of the system, results were examined with a high-level perspective. We applied an additional logical rule on top of our classifier to correct some misclassifications. With decision level smoothing, the classification accuracies increased to around 90 % with both modalities (92.89% maximum). We further developed a session-based stress classifier. Majority voting among windows of every session was applied to decide the assigned class. We obtained a maximum accuracy of 94.44% with HR, 100% with EDA signals. When the stress level of a session is needed to be calculated, this method could be applied.

We improved our platform independent stress level detection system which works with off-the-shelf smartwatches and smart bands. We tested our algorithms in a real-life setting and obtained successful classification accuracies. As mentioned, personal stress level clustering and decision-level smoothing increased the performance of our system considerably. We also applied stress alleviation methods and proved their effectiveness. Our system could be easily adapted to daily life of individuals without interrupting their routines. The study has limitations that should be mentioned. With regard to the measurement, we have initially included NASA-TLX which is a cognitive workload scale. In order to measure the perceived stress levels, as explained in the methods section, we have selected the frustration subscale which is the most representative for that purpose.

Nevertheless, future studies specifically focusing on the perceived stress could better include specific scales such as Daily Stress Inventory [57], Daily Experiences Survey [58] or Perceived Stress Scale [59]. As a future study, we plan to develop personalized perceived stress models from self-reports.

REFERENCES

[1] Y. S. Can, B. Arnrich, and C. Ersoy, "Stress detection in daily life scenarios using smart phones and wearable sensors: A survey," *J. Biomed. Informat.*, vol. 92, Apr. 2019, Art. no. 103139.

[2] *Generalized Anxiety Disorder*. Accessed: Feb. 2020. [Online]. Available: https://www.webmd.com/balance/stress-management/qa/what-are-the-consequences-of-longterm-stress

[3] P. Cheng, A. Lucero, and J. Buur, "Pause: Exploring mindful touch interaction on smartphones," in *Proc. 20th Int. Acad. Mindtrek Conf.* New York, NY, USA: Academic, 2016, pp. 184–191.

[4] R. W. Picard, "Automating the recognition of stress and emotion: From lab to real-world impact," *IEEE Multimedia Mag.*, vol. 23, no. 3, pp. 3–7, Jul. 2016.

[5] Y. S. Can, N. Chalabianloo, D. Ekiz, and C. Ersoy, "Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study," *Sensors*, vol. 19, no. 8, p. 1849, Apr. 2019.

[6] E. Vildjiounaite, J. Kallio, V. Kyllönen, M. Nieminen, I. Määttänen, M. Lindholm, J. Mäntyjärvi, and G. Gimel'farb, "Unobtrusive stress detection on the basis of smartphone usage data," *Pers. Ubiquitous Comput.*, vol. 22, no. 4, pp. 671–688, Jan. 2018.

[7] R. Castaldo, L. Montesinos, P. Melillo, S. Massaro, and L. Pecchia, "To what extent can we shorten HRV analysis in wearable sensing? A case study on mental stress detection," in *EMBEC & NBC*, H. Eskola, O. Väisänen, J. Viik, and J. Hyttinen, Eds. Singapore: Springer, 2018, pp. 643–646.

[8] J. R. Machado Fernández and L. Anishchenko, "Mental stress detection using bioradar respiratory signals," *Biomed. Signal Process. Control*, vol. 43, pp. 244–249, May 2018.

[9] G. Giannakakis, M. Pediaditis, D. Manousos, E. Kazantzaki, F. Chiarugi, P. G. Simos, K. Marias, and M. Tsiknakis, "Stress and anxiety detection using facial cues from videos," *Biomed. Signal Process. Control*, vol. 31, pp. 89–101, Jan. 2017.

[10] O. M. Mozos, V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, R. Dobrescu, and J. M. Ferrandez, "Stress detection using wearable physiological and sociometric sensors," *Int. J. Neural Syst.*, vol. 27, no. 2, Dec. 2016, Art. no. 1650041.

[11] R. Martinez, E. Irigoyen, A. Arruti, J. I. Martin, and J. Muguerza, "A real-time stress classification system based on arousal analysis of the nervous system by an F-state machine," *Comput. Methods Programs Biomed.*, vol. 148, pp. 81–90, Sep. 2017.

[12] B. Egilmez, E. Poyraz, W. Zhou, G. Memik, P. Dinda, and N. Alshurafa, "UStress: Understanding college student subjective stress using wrist-based passive sensing," in *Proc. IEEE Int. Conf. Pervas. Comput. Commun. Workshops (PerCom Workshops)*, Mar. 2017, pp. 673–678.

[13] M. Ciman and K. Wac, "Individuals' stress assessment using human-smartphone interaction analysis," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 51–65, Jan. 2018.

[14] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "Continuous stress detection using a wrist device: In laboratory and real life," in *Proc. ACM Int. Joint Conf. Pervas. Ubiquitous Comput., Adjunct (UbiComp)*, New York, NY, USA, 2016, pp. 1185–1193.

[15] M. Sysoev, A. Kos, and M. Pogăčnik, "Noninvasive stress recognition considering the current activity," *Pers. Ubiquitous Comput.*, vol. 19, no. 7, pp. 1045–1052, Aug. 2015.

[16] A. Muaremi, A. Bexheti, F. Gravenhorst, B. Arnrich, and G. Troster, "Monitoring the impact of stress on the sleep patterns of pilgrims using wearable sensors," in *Proc. IEEE-EMBS Int. Conf. Biomed. Health Inform. (BHI)*, Jun. 2014, pp. 185–188.

[17] A. Sano and R. W. Picard, "Stress recognition using wearable sensors and mobile phones," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 671–676.

[18] K. Hovsepian, M. Al'Absi, E. Ertin, T. Kamarck, M. Nakajima, and S. Kumar, "cStress: Towards a gold standard for continuous stress assessment in the mobile environment," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. (UbiComp)*, New York, NY, USA, 2015, pp. 493–504.

IEEE Access

Y. S. Can *et al.*: Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory

[19] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," *J. Biomed. Informat.*, vol. 73, pp. 159–170, Sep. 2017.

[20] A. de Santos Sierra, C. S. Avila, J. G. Casanova, G. B. del Pozo, and V. J. Vera, "Two stress detection schemes based on physiological signals for real-time applications," in *Proc. 6th Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, Oct. 2010, pp. 364–367.

[21] V. Sandulescu, S. Andrews, D. Ellis, N. Bellotto, and O. M. Mozos, "Stress detection using wearable physiological sensors," in *Proc. Int. Work-Conf. Interplay Between Natural Artif. Comput.* Cham, Switzerland: Springer, 2015, pp. 526–532.

[22] V. Vanitha and P. Krishnan, "Real time stress detection system based on EEG signals," *Biomed. Res.*, pp. 271–275, 2016.

[23] J. Aigrain, S. Dubuisson, M. Detyniecki, and M. Chetouani, "Person-specific behavioural features for automatic stress detection," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, vol. 3, May 2015, pp. 1–6.

[24] J. Aigrain, M. Spodenkiewicz, S. Dubuisson, M. Detyniecki, D. Cohen, and M. Chetouani, "Multimodal stress detection from multiple assessments," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 491–506, Oct. 2018.

[25] J.-H. Hong, J. Ramos, and A. K. Dey, "Understanding physiological responses to stressors during physical activity," in *Proc. ACM Conf. Ubiquitous Comput. (UbiComp)*, New York, NY, USA, 2012, pp. 270–279.

[26] A. O. Akmandor and N. K. Jha, "Keep the stress away with SoDA: Stress detection and alleviation system," *IEEE Trans. Multi-Scale Comput. Syst.*, vol. 3, no. 4, pp. 269–282, Oct. 2017.

[27] M. A. B. S. Akhonda, S. M. F. Islam, A. S. Khan, F. Ahmed, and M. M. Rahman, "Stress detection of computer user in office like working environment using neural network," in *Proc. 17th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2014, pp. 174–179.

[28] M. N. Haji Mohd, M. Kashima, K. Sato, and M. Watanabe, "Mental stress recognition based on non-invasive and non-contact measurement from stereo thermal and visible sensors," *Int. J. Affect. Eng.*, vol. 14, no. 1, pp. 9–17, 2015.

[29] A. Liapis, C. Katsanos, D. Sotiropoulos, M. Xenos, and N. Karousos, "Stress recognition in human-computer interaction using physiological and self-reported data: A study of gender differences," in *Proc. 19th Panhellenic Conf. Inform. (PCI)*, New York, NY, USA, 2015, pp. 323–328.

[30] E. Garcia-Ceja, V. Osmani, and O. Mayora, "Automatic stress detection in working environments from Smartphones' accelerometer data: A first step," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 4, pp. 1053–1060, Jul. 2016.

[31] B. Cinaz, B. Arnrich, R. La Marca, and G. Tröster, "Monitoring of mental workload levels during an everyday life office-work scenario," *Pers. Ubiquitous Comput.*, vol. 17, no. 2, pp. 229–239, Oct. 2011.

[32] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 156–166, Jun. 2005.

[33] M. Gjoreski, H. Gjoreski, M. Lutrek, and M. Gams, "Automatic detection of perceived stress in campus students using smartphones," in *Proc. Int. Conf. Intell. Environ.*, Jul. 2015, pp. 132–135.

[34] A. Bogomolov, B. Lepri, M. Ferron, F. Pianesi, and A. S. Pentland, "Pervasive stress recognition for sustainable living," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PERCOM WORKSHOPS)*, Mar. 2014, pp. 345–350.

[35] G. Bauer and P. Lukowicz, "Can smartphones detect stress-related changes in the behaviour of individuals?" in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops*, Mar. 2012, pp. 423–426.

[36] K. Chen, W. Fink, J. Roveda, R. D. Lane, J. Allen, and J. Vanuk, "Wearable sensor based stress management using integrated respiratory and ECG waveforms," in *Proc. IEEE 12th Int. Conf. Wearable Implant. Body Sensor Netw. (BSN)*, Jun. 2015, pp. 1–6.

[37] S. Taylor, N. Jaques, W. Chen, S. Fedor, A. Sano, and R. Picard, "Automatic identification of artifacts in electrodermal activity data," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 1934–1937.

[38] A. Greco, G. Valenza, A. Lanata, E. Scilingo, and L. Citi, "CvxEDA: A convex optimization approach to electrodermal activity processing," *IEEE Trans. Biomed. Eng.*, vol. 63, pp. 797–804, Apr. 2016.

[39] A. Alberdi, A. Aztiria, and A. Basarab, "Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review," *J. Biomed. Informat.*, vol. 59, pp. 49–75, Feb. 2016.

[40] S. Greene, H. Thapliyal, and A. Caban-Holt, "A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health," *IEEE Consum. Electron. Mag.*, vol. 5, no. 4, pp. 44–56, Oct. 2016.

[41] C. Kappeler-Setz, *Multimodal Emotion and Stress Recognition*. Zürich, Switzerland: ETH Zürich, 2012.

[42] I. V. Bornoiu and O. Grigore, "A study about feature extraction for stress detection," in *Proc. 8th Int. Symp. Adv. Topics Electr. Eng. (ATEE)*, May 2013, pp. 1–4.

[43] M. P. Tarvainen, J.-P. Niskanen, J. A. Lipponen, P. O. Ranta-Aho, and P. A. Karjalainen, "Kubios HRV–heart rate variability analysis software," *Comput. Methods Programs Biomed.*, vol. 113, no. 1, pp. 210–220, 2014.

[44] M. Vollmer, *MarcusVollmer/HRV Toolbox*. San Francisco, CA, USA: GitHub, 2019.

[45] N. R. Lomb, "Least-squares frequency analysis of unequally spaced data," *Astrophys. Space Sci.*, vol. 39, no. 2, pp. 447–462, Feb. 1976.

[46] *Stress Response*. Accessed: Feb. 2020. [Online]. Available: https://www.anxietycentre.com/anxiety/stress-response.shtml

[47] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[48] F. Eibe, M. Hall, and I. Witten, "The WEKA workbench. Online appendix for," in *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2016.

[49] S. G. Hart, "NASA-task load index (NASA-TLX); 20 years later," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 50. Los Angeles, CA, USA: Sage, 2006, pp. 904–908.

[50] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in *Advances in Psychology*, vol. 52, P. A. Hancock and N. Meshkati, Eds. Amsterdam, The Netherlands: North Holland, 1988, pp. 139–183.

[51] E. Galy, M. Cariou, and C. Mélan, "What is the relationship between mental workload factors and cognitive load types?" *Int. J. Psychophysiol.*, vol. 83, no. 3, pp. 269–275, 2012.

[52] G. D. Rey and F. Buchwald, "The expertise reversal effect: Cognitive load and motivational explanations.," *J. Exp. Psychol., Appl.*, vol. 17, no. 1, pp. 33–48, Mar. 2011.

[53] S. G. Hart, "Nasa-task load index (NASA-TLX); 20 years later," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, Nov. 2016, vol. 50, no. 9, pp. 904–908.

[54] *HeartMath App*. Accessed: Feb. 2020. [Online]. Available: https://store.heartmath.com/inner-balance/

[55] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Troster, and U. Ehlert, "Discriminating stress from cognitive load using a wearable EDA device," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 410–417, Mar. 2010.

[56] B. Arnrich, C. Setz, R. La Marca, G. Troster, and U. Ehlert, "What does your chair know about your stress level?" *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 207–214, Mar. 2010.

[57] P. J. Brantley and G. N. Jones, *Daily Stress Inventory: Professional Manual*. Lake Magdalene, FL, USA: Psychological Assessment Resources Odessa, 1989.

[58] S. R. Stader and J. E. Hokanson, "Psychosocial antecedents of depressive symptoms: An evaluation using daily experiences methodology," *J. Abnormal Psychol.*, vol. 107, no. 1, pp. 17–26, 1998.

[59] S. Cohen, T. Kamarck, and R. Mermelstein, "A global measure of perceived stress," *J. Health Social Behav.*, vol. 24, no. 4, pp. 385–396, Dec. 1983.

YEKTA SAID CAN received the B.Sc. degree in computer engineering from the Faculty of Engineering, Boğaziçi University, Istanbul, Turkey, in 2012, and the M.Sc. degree from Boğaziçi University, in 2014, where he is currently pursuing the Ph.D. degree in computer engineering. He worked as a Researcher at TUBITAK BILGEM for two years. He also works as a Teaching Assistant with the Faculty of Engineering, Boğaziçi University. His research interests include watermarking, speech and speaker recognition, physiological signal processing, and machine learning.

Y. S. Can *et al.*: Personal Stress-Level Clustering and Decision-Level Smoothing to Enhance the Performance of Ambulatory

IEEE *Access*

NIAZ CHALABIANLOO received the master's degree in computer engineering from Middle East Technical University, Turkey. He is currently pursuing the Ph.D. degree with Boğaziçi University, where he is working on multiparametric monitoring of mood and sleep quality for stress-management. He is also an Early Stage Researcher for the Marie Sklodowska-Curie H2020 AffecTech Project. His current research interests include data science, pervasive health, ubiquitous computing, and wireless sensor networks.

DENIZ EKIZ is currently pursuing the Ph.D. degree with the Computer Engineering Department, Boğaziçi University. His research is focused on the health-related applications of wearable technology.

JAVIER FERNANDEZ-ALVAREZ is currently pursuing the Ph.D. degree with the Universita Cattolica del Sacre Cuore. He does research in clinical psychology and psychotherapy using new digital technologies. He takes part of AffecTech Project.

GIUSEPPE RIVA is currently a Full Professor of general psychology with the Universita Cattolica del Sacre Cuore and also the Director of the Applied Technology for Neuro-Psychology Laboratory, Istituto Auxologico Italiano. He is a member of the American Psychological Association. He is actually an Editor in Chief for the *Emerging Communication* book series and an European Editor of the *CyberPsychology, Behavior & Social Networking* scientific journal.

CEM ERSOY (Senior Member, IEEE) received the Ph.D. degree from Polytechnic University, New York, NY, USA, in 1992. He was an Research and Development Engineer with NETAS A. S., from 1984 to 1986. He is currently a Professor of computer engineering with Boğaziçi University. He is also the Vice Director of the Telecommunications and Informatics Technologies Research Center, TETAM. His research interests include wireless/cellular/ ad-hoc/sensor networks, activity recognition, and ambient intelligence for pervasive health applications, green 5G and beyond networks, mobile cloud/edge/fog computing, software-defined networking, and infrastructure-less communications for disaster management. He is a member of IFIP. He was the Chairman of the IEEE Communications Society Turkish Chapter during the last 7 years.

● ● ●