UStress: Understanding College Student Subjective Stress Using Wrist-Based Passive Sensing

Begum Egilmez*, Emirhan Poyraz*, Wenting Zhou*, Gokhan Memik*, Peter Dinda*, Nabil Alshurafa†

*Dept. of Electrical Eng. and Computer Science
Northwestern University
Evanston, IL, USA

†Department of Preventive Medicine
Northwestern University
Chicago, IL, USA

Abstract—Stress plays a major role in physical and emotional well-being, and is associated with several illnesses including depression, diabetes, and other chronic diseases. College student stress as a construct is important to detect in order to equip students in a timely manner with stress coping strategies. However, the lack of a passive sensing measure accepted as a gold standard impedes real time detection and treatment of stress. Many researchers are studying passive sensing of stress using wrist-worn sensors; however, this effort focuses on understanding the essential features of wrist-worn sensors in detecting stress, and how to best induce stress in a lab setting.

Applying machine learning methods increasingly is making it feasible to validly infer in real time through passive sensing of physical features psychological states, such as stress. Given strong participant adherence to wrist-worn sensors, this paper focuses on analyzing the effect of replacing other body sensing platforms (e.g. chest-based heart rate) with their wrist-worn equivalent on stress prediction accuracy. Nine participants were equipped with multiple body sensors and were asked to wear a commercially available Android smartwatch, a custom-designed smartwatch equipped with a Galvanic Skin Response (GSR) sensor, a chest-based heart rate sensor, and a finger-based commercial GSR sensor. Based on participant self-reports, singing experiment showed greatest stress levels across participants. This paper further analyzes features for prediction from all sensors compared to wrist-worn only sensors. Using statistical features on one-minute fixed-time sub-divisions and correlation-based feature subset selection and a Random Forest model, the system is capable of detecting stress with 88.8% F-measure.

Index Terms—Android, smartwatch, wrist sensors, stress detection, passive sensing, heart rate, galvanic skin response, machine learning.

I. INTRODUCTION

Stress can be a positive force, encouraging students to study more to learn and produce knowledge, and other times it is harmful and may lead to psychological and physiological diseases. As a direct or indirect source of many health issues, stress is inclined to lower the individual’s life quality and contributes to increasing number of suicide cases. Studies show one in every four Americans exhibit great levels of stress [14]. In a recent survey of 120 people, we show that college students primarily feel stress from: Homework/Projects, Grades, and Exams. Another recent survey shows that around 49% of students report feeling stress on a daily basis, particularly females [15].

The StudentLife [24] project lays the groundwork for stress and passive sensing in college students, showing the need to study stress, and how it varies throughout the quarter, positively correlating with sleep and conversation duration and frequency. Recent passive sensing stress research has shown success of combining a wrist-worn sensor with a chest-based ECG sensor to detect stress based on models built in-lab [19][10]. Many researchers are also designing new devices and applying machine learning models in the lab. However, to do so, they must start by testing in a lab setting. Despite the increasing research on detecting stress, it is not well understood which features are most important in detecting subjective stress in a model generalized purely from data collected from wrist-worn sensors. Moreover, it is not well understood which activities successfully induce stress on college students in a lab setting.

The ability to passively detect stress in lagged real-time can enable interventions that help students cope with stress. Such a mechanism can help support student feedback to understand their stress patterns (time, frequency, and duration) and test interventions to understand and ultimately treat college student stress.

Previous studies have shown that there are multiple physiological markers that can be affected by the presence of cognitive stress. These markers are related to heart rate, GSR, blood pressure, and respiration rate. Since heart rate and GSR can currently be reliably sensed passively in existing wrist-worn wearables [6][12], this paper focuses on using these sensors from a commercially available smartwatch. Inertial sensors are also used to capture predictive features of stress and also to potentially filter false alarms due to physical activity.

This work contributes to understanding subjective stress by:

- identifying the most successful stress induction methods in college students;
- defining predictive features to detect subjective stress in college students from wrist- and chest-worn sensors; and
- defining predictive features to detect subjective stress in college students only from wrist-worn sensors.

2https://afsp.org/about-suicide/suicide-statistics/
To this end, first, the state-of-the-art technology in stress detection is discussed and compared to that of this work. The following section describes system details including sensing devices, data collection, and experimental setup. Followed by the methodology section which explains how the system is trained and tested, and finally, results are discussed, ending with conclusions and opportunities for future work.

II. RELATED WORK

StressSense [13] measures stress from human voice using a smartphone microphone, recognizing changes in speech such as pitch range, jitter and speaking rate. This system implements indoor-outdoor speech-based experiments, reporting 80% accuracy of stress prediction. One factor that limits the use of this technology is the need for the user to speak, limiting its usage to detecting stress in vocal participants, failing to detect stress when people are silent (e.g. a college student doing homework, or taking an exam). This system is also affected by noise from the environment.

To overcome this problem, recent passive sensing techniques show promise using continuous measurement of physiological parameters such as heart rate and galvanic skin response (GSR). Vrijkotte et. al. [22] detects work-life stress using wearable sensors that capture ambulatory blood pressure, heart rate and heart rate variability. BeWell [11] and StudentLife [24] are Android applications to assess the stress level of the smartphone user by tracking activities that affect physical, social and mental well-being. The relevant data is collected by continuous measurement of smartphone embedded sensors including the microphone, accelerometer, and light sensor. Due to increased likelihood of participant adherence to wearing a wrist-worn sensor (compared to other wearables), using a wrist-worn sensor further expands the ability to continuously sense physiological parameters, enabling reliable passive sensing of stress.

Commercial smartwatches are currently equipped with common sensors that detect body temperature, heart rate, and accelerometer and gyroscope sensors, however, many are not equipped with GSR sensing. The Embrace watch [5] is one commercial wrist-worn sensor that measures stress, tracks activity, and monitors sleep, but does not show time. The Envio [7] is slated to measure stress only by heart rate variability, but is not available for purchase. These watches do not target college student stress specifically, and may benefit from expanding their feature set to detect college student stress.

Researchers are increasingly building their own stress detection platforms based on their own needs. Ouwerkerk et. al. [16] develops a bracelet with built-in sensors including GSR, accelerometer, device temperature, and ambient light level. Hovsepian et. al. [10] creates a stress model (cStress) to standardize every step of computational modeling including data collection, screening, cleaning, filtering, feature computation, normalization, and model training. In this study, they collect data from multiple body sensors, focusing on ECG chest-based data, to test the models built in lab on models in the field. This paper attempts to build upon this related work by uncovering the predictive features of wrist-worn sensors in predicting subjective stress in college students.

III. SYSTEM

The objective of this paper is to induce stress in a controlled environment and identify the physiological responses that can help accurately predict stress using features only from wrist-worn sensor data. The particular system design, as well as the sensors and devices used during the experiments, capture participants’ physiological responses to stress.

A. Stress-Inducing Experiment Design

In these stress induction experiments, participants are invited to the lab, and requested to wear sensors and perform different tasks. None of the participants knew what tasks they would perform until they came to the lab. Upon arrival they wore the sensors, seated in front of a PC and were requested to comply with the tasks displayed on the screen. They were allowed to terminate the test anytime they felt uncomfortable, but were not allowed to skip tasks.

Table I shows the details of tasks performed. The stress-based tasks were randomized for each participant to minimize carry-over effects. All these tasks had a fixed duration of 4 minutes, except the ice-bucket test, where participants were encouraged to keep their hand in ice for as long as possible. At the end of each task, participants were asked to rate their stress level on a scale of 1 to 5 [2], 1 being the lowest and 5 the highest stress they face in daily life. No definition of stress was provided to capture subjective stress. After they submitted their rating, a 2-minute rest interval allowed participants to calm down and get ready for the next task. Regular tasks such as eating, engaging in conversation, and doing homework were interspersed with stressful activities. Each activity is represented by S (stress) or NS (no stress) class label. The reason for incorporating NS tasks is to create a test environment that includes real-life tasks which may create mild stress, and also challenges the prediction model. We did not include confounding physical activity-based tasks since that could be detected and filtered from an accelerometer in future studies. Fig. 1 shows an example flow of tasks with photos from different subjects’ tests.

B. Devices and Sensors

It has been shown that there is a relationship between stress and skin conductivity or galvanic skin response (GSR) [21]. This study uses a custom-built, portable wrist-worn GSR device (WGSR) that can be seen in Fig. 2, worn on the left wrist, next to an LG smartwatch. WGSR is based on the Northwestern-developed NUSensor platform, which eases the board- and software-level design and implementation of a small rechargeable device that can sample data directly from a set of sensor ICs and analog sensors at high rates, transmit it to an iOS or Android phone via BLE, and decode, display, and transmit the data to a smartphone. To validate WGSR, a commercially available, but non-portable GSR sensor was also used, namely a NeuLog by Eisco.
Fig. 1: Example flow of activities in a test. Photos representing each activity are taken from real in-lab tests.

Labs [4], which we refer to as NGSR. Both NGSR and WGSR have 10-bit ADC and 100ns resolution. Note that NGSR is not a mobile device and must be connected to a PC to transmit data. In future work we intend to study stress outside of the lab environment, hence the use of WGSR.

Apart from skin response, heart rate is another indicator of stress in the human body. Since most of the commercially available smartwatches in the market are equipped with heart rate and many other sensors, participants also wore an LG Watch Urbane 2. Given the recent concern of heart-rate accuracy from wrist-worn devices [8], however, a chest-worn heart rate sensor, Polar H7 [17], is also used to test the reliability of the wrist-worn sensor.

The tri-axial accelerometer and gyroscope data embedded into the LG smartwatch are processed as well, because people move in unique ways under stress. This data also helps distinguish between high and low-intensity activity.

All the sensors used are shown in Figure 2. Since participants wore the Polar H7 chest-band under clothing for direct skin contact, the right image shows how to position the Polar sensor on the body. The study coordinator aided participants in wearing the other devices so that the devices were contacting the skin and comfortable to wear (i.e not tight or loose).

### TABLE I: Activities Included

<table>
<thead>
<tr>
<th>Activity</th>
<th>Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ice Bucket</td>
<td>S</td>
<td>Dunk dominant hand into ice water for as long as possible.</td>
</tr>
<tr>
<td>Singing</td>
<td>S</td>
<td>Sing songs out loud without any background music.</td>
</tr>
<tr>
<td>Game</td>
<td>S</td>
<td>Play the game &quot;The Case of Scary Shadows.&quot;</td>
</tr>
<tr>
<td>Stroop</td>
<td>S</td>
<td>Type in the font color of text appearing on the screen, which reads the name of a different color. Participants should try scoring as many correct answers as possible before time runs out.</td>
</tr>
<tr>
<td>Math</td>
<td>S</td>
<td>Answer as many arithmetic questions as possible before running out of time.</td>
</tr>
<tr>
<td>SocCon</td>
<td>NS</td>
<td>Engage in light conversation with researcher in charge.</td>
</tr>
<tr>
<td>Homework</td>
<td>NS</td>
<td>Do homework.</td>
</tr>
<tr>
<td>Emails</td>
<td>NS</td>
<td>Read, write, or reply to emails.</td>
</tr>
<tr>
<td>Eating</td>
<td>NS</td>
<td>Eat one of the complimentary snack options.</td>
</tr>
</tbody>
</table>

### C. Data Collection

WGSR transmits data to its paired Android smartphone via Bluetooth. NGSR has its own software for logging data. Similar to WGSR, the LG smartwatch has a corresponding custom Android application to communicate its readings, with the capability to transmit to a backend server. The chest-band Polar H7 transmits its readings to an Android-based Pulsometer RR app. This application logs the readings in a file and stores it in local memory. All data is collected at a 5Hz sampling rate, except the Polar H7, which samples at 1Hz. However, to synchronize with other data streams it is upsampled to 5Hz.

An example of raw data collected from the heart rate sensors (smartwatch and Polar H7) and the GSR sensors (WGSR and NGSR) is shown in Figure 3. GSR readings from both sensors do not always match one-to-one, but capture the overall trend. The interval marked with red arrows in Figure 3 is where the WGSR sensor provided incorrect readings, which may be the result of loosened probes. The fact that these misreadings begin in the Eating activity supports this theory, because most of the participants unintentionally want to use both hands while eating. Heart rate readings from Polar and LG watch also show correlated trend in data.

### IV. METHODOLOGY

This section provides details of the methodology used to predict subjective stress.

Fig. 3: GSR and heart rate data collected from one participant. R1 to R7 refer to rest. Other activities performed were: Stroop, Game, Math, Eating (Eat), Singing (Sing), Homework (HW), Emails, Social Conversation (SocCon) and Ice. Interval marked with red arrow shows poor quality data.

A. Feature Extraction

Among 9 study participants, 7 provided reliable, complete data. One of the other two had significant amount of sensor readings missing and other one did not complete all assigned tasks.

Features extracted on each time-series data stream are described in Table II, with 110 total features. All features are based on statistical analysis of the data across the entire activity performed, to enable real-time and less-intensive computations. Two approaches for feature-extraction are used, Event-Based (EB), and Minute-Based (MB). The EB approach assumes that the beginning and end of the task are known, and a window surrounding the start and end of each activity is used to calculate statistical features across the entire window. The MB approach is designed for a lag real-time prediction, using a window-size of one minute, with 75% overlap between windows.

### TABLE II: Features Extracted from Sensors

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>Maximum value in the selected window</td>
</tr>
<tr>
<td>Min</td>
<td>Minimum value in the selected window</td>
</tr>
<tr>
<td>Mean</td>
<td>Average of values in the selected window</td>
</tr>
<tr>
<td>Median</td>
<td>Median of values in the selected window</td>
</tr>
<tr>
<td>StDev</td>
<td>Standard deviation of values in the selected window</td>
</tr>
<tr>
<td>Skew</td>
<td>Symmetry of distribution of values in the window</td>
</tr>
<tr>
<td>RMS</td>
<td>Root-mean-square of values in the selected window</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>Measure for outliers in the selected window</td>
</tr>
<tr>
<td>Quart1</td>
<td>Median of lower 25% of the values in the window</td>
</tr>
<tr>
<td>Quart3</td>
<td>Median of upper 25% of the values in the window</td>
</tr>
<tr>
<td>IRQ</td>
<td>Difference between Quart3 and Quart1</td>
</tr>
</tbody>
</table>

B. Feature Selection

To avoid the problem of overfitting and to produce more interpretable predictive models, a feature selection algorithm is used to alleviate the effects of feature redundancy while increasing predictive power. Correlation-based Feature Subset Evaluation (CfsSubset) is used [23] to evaluate the importance of a subset of attributes by considering their individual prediction ability and degree of redundancy between them.

C. Classification Approach

Two models are developed here. The first model is the intended-stress model (labelled i), which is used for predicting the intended-stress outcome variable, according to the Label column in Table I. Because stress is experienced variably across people, a second model is developed called the self-reported stress model (labelled s), which is used to predict the self-reported subjective stress level. As each individual exhibits a different stress threshold, mean of all activities is calculated for each individual and used to distinguish a likely stress (S) activity from a non-stress activity (NS). For example, one participant provided a mean rating of 2.56 for all 9 activities, so any rating below this number is labeled as NS and anything above is labeled S.

Both intended-stress and self-reported stress models are designed for all the sensors combined and the wrist-only sensors, in order to understand if features held to be important when using the combined sensors transfer to the wrist-only model. If they are both truly measuring the same heart accurately, the features would remain consistent across body position. Different algorithms used in prior stress-based research were used to develop our models including NaiveBayes, SVM, Logistic Regression, and Random Forest. Each of these algorithms are tested both with Leave One Subject Out Cross Validation (LOSOcv) and 10-fold Cross Validation.

V. RESULTS

This section discusses the predictive features, performance of stress predictive models, and the effectiveness of in-lab stress induction.

A. Selected Features

Table III lists the features selected for each dataset. Dataset names depend on three factors: the windows used to extract the features (EB or MB), the type of sensor (WC: all wrist and chest sensors; W: wrist-only sensors), and labeling approach (i: intended-stress; s: self-reported stress). The Features column displays the statistical features extracted. Here, WHR refers to the wrist-worn heart-rate sensor in the LG smartwatch, whereas CHR indicates the heart rate sensor from the chest-worn Polar H7. Similarly, NGSR is the GSR sensor of NeuLog, and WGSR refers to the wrist-worn custom-built GSR. GYRY is the y-axis gyroscope value read from LG watch.

Interestingly, CHR is not proven to be a predictive feature, but WHR is included as such. Replacing features collected from the wrist-worn heart rate sensor with those of the...
chest-worn heart rate sensor only improves the f-measure by 1%. This finding highlights a potential problem with feature selection algorithms, because they do not exhaustively search all possible options, they may not capture the global optimal set of features.

Although CfsSubset generates a subset of important features for MB-WCs and MB-Ws models, classification performed using those features results in poor accuracy (61.7%, 70%, respectively). As a result, combining CfsSubset with a wrapper-based method [9] using the Random Forest algorithm as the classifier produces more reliable results (discussed in Section V.C.2.), selecting 37 predictive features that are predominantly selected from WHR, NGSR, WGSR, and gyroscope (GYRX, GYRY, GYRZ) data.

TABLE III: Features Selected for Each Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>EB-WCi</td>
<td>NGSR_mean, WHR_kurtosis</td>
</tr>
<tr>
<td>EB-WCs</td>
<td>NGSR_mean, NGSR_skew, WHR_min,</td>
</tr>
<tr>
<td></td>
<td>GYRY_kurtosis, WGRS_irq</td>
</tr>
<tr>
<td>EB-Ws</td>
<td>WHR_min, GYRY_kurtosis, WGRS_irq,</td>
</tr>
<tr>
<td></td>
<td>WHR_mean</td>
</tr>
</tbody>
</table>

B. Classifier and Dataset Accuracies

1) Best classifier: For both EB-WCi and EB-WCs datasets, several classifiers were tested, including: Naive-Bayes (NB), SVM, Logistic Regression (LR) and Random Forest (RF). The results for EB-WCi (intended stress) were: 59.2% NB, 58.4% SVM, 57.6% LR, and 59.1% RF. The results for EB-WCs (self-reported stress) were: 26.8% NB, 61.3% SVM, 66.1% LR, and 78.8% RF. All classifiers yield significant improvement in F-measure for reported stress compared to intended stress labels. This is not surprising given that subjective stress differs from one subject to another. This result shows the importance of a personalized stress induction methodology. Since Random Forest outperforms other classifiers it is used in the remainder of our analysis.

2) Dataset Accuracies with Random Forest Model: As shown in Table II, features selected for EB-WCi are just NGSR_mean and WHR_kurtosis, highlighting the importance of skin conductivity and heart-rate as important indicators of stress. When the selected NGSR feature is replaced with the custom-built WGSR version, all evaluation metrics of the model improve slightly both in predicting intended stress and in self-reported stress.

Besides these two feature sets, the EB-Ws dataset has a predictive feature set similar to EB-WCs, and when using the features extracted from EB-WCs on EB-Ws (but using their wrist-based counterpart), there is a negligible (1.7%) decrease in F-measure. Also, using the features collected in EB-Ws on EB-WCs produces similar results. This shows that there are multiple features that could be used to distinguish levels of stress (as seen in Figures 4 and 5). Such proximity in the results show that wrist-only based versions of physiological sensors can still provide reliable stress predictions, even though their readings do not precisely match, as the correlation between their ground truth counterparts is significant enough to produce useful features.

Applying a wrapper-based feature selection [9] method (with the Random Forest classifier) on the last two datasets, MB-WCs and MB-Ws, results in 37 predictive features. The features found predictive in the MB-WCs dataset are also found predictive in their alternative sensor counterpart in the MB-Ws dataset. The evaluation metrics for these two datasets are shown in Figure 6, and outperform those of the EB models. The C. Kappa statistics of the WC model and W model are 0.63 and 0.67, respectively. Results therefore support the potential to use wrist-only sensors to predict stress in college students, with real-time stress also showing great potential.

C. Best Stress-Inducing Activity

Although in-lab stress induction can be a challenging task, participant self-reported stress levels show great variability in how people respond to stress. Figure 7 shows the singing stress test is inducing the greatest stress in college students, followed by Math test (in the presence of someone watching). On the opposite end of the scale, eating is shown to exhibit the least amount of stress, which justifies why people use...
validate the usability and accuracy of our sensing platform in stress detection.

REFERENCES


V. CONCLUSION AND FUTURE WORK

VI. CONCLUSION AND FUTURE WORK

Varying stress inducing experiments are performed in-lab on 9 college students using chest-worn, finger-worn and wrist-worn wearable devices. Both coarse-grained and fine-grained analysis show that even though the readings of wrist- and chest-worn heart rate sensors do not exhibit strong cross correlation, they can be used interchangeably for stress detection. For event-based classification, 78.8% f-measure is achieved when NGSR_mean, NGSR_skew, WHR_min, GYR_kurtosis, and WGSR_irq are used for prediction. For minute-based classification, f-measure is 88.8%. The classifier comparisons show that Random Forest is the best performing model for the data collected in this work. Finally, this work shows that it is possible to trigger different levels of stress in-lab, with the singing test showing the greatest level of subjective stress among college students. Future work stemming from this study includes developing an accurate personalized real-time stress detection model based solely on wrist-worn sensors, and testing its performance in the wild with a larger group of subjects. Since this effort is based on a controlled lab environment, future in-the-wild studies will