Mobile Sensors and Health: Predicting Stress from Physiological Data

Kevin Sun, Megan Stiles, Matt Davolio, Hampton Leonard

Abstract

The rising use of personal health and fitness measurement tools have allowed individuals to record and track many aspects of their lives. However, much of this data being recorded is being underutilized in its current state and simply being observed rather than analyzed. In order to bring further value to this collected data, it is thought to be possible to relate these recorded factors to user's psychological state. A state which is of high interest is if the user is stressed. If a user can be warned that they may become stressed in the near future based on various reading from a personal health tool, steps can be taken for the individual to attempt to avoid this state. Using data collected from such a tool will be utilized to attempt to predict a user's response to a survey related to their current stress level. Largely due to the design and optional participation of the experiment, it was found that many of the different readings recorded by health trackers are not indicative of an individual's upcoming stress level.

Descriptive Analysis

Stress is a common problem in society that comes with significant costs. In the 2000 European Working conditions Survey, work-related stress was identified as the second most common work-related health problem across the EU (Paoli). By some estimates, work-related stress costs US businesses $30 billion from lost workdays each year.[1] According to a 2004 APA survey, 25% of employees have taken a day off work in order to deal with the side effects of stress.[2] Stress can also incur costs on an individual level. When stressed, consumers often respond by making excessive expenditures on what they consider "necessities". These can be items such as household goods, but they can also be items that are detrimental to health, such as high-caloric foods.[3] While increased spending is good for the market, a lack of understanding of this bias towards spending increased amounts on certain items when stressed can cause huge overall costs to individuals.[3] Understanding the causes of stress can help to avoid monetary losses on both the corporate and individual level.

Stress can lead to illness both directly and indirectly. Stress causes physiological reactions in the body by affecting both the nervous system and the endocrine system, which control a human's "flight or fight" mechanism and the production of hormones such as adrenaline and cortisol, respectively.[4] The responses to these systems being activated include increased heart rate, increased blood pressure, activation of sweat secretion, and release of glucose and fats from storage sites into the body, among others.[4] While often necessary for survival in threatening situations, prolonged exposure to these stress responses have been proven to have negative effects on overall health. Stress impairs working memory and cognitive

---

[1] Pazzanese, Christina "The High Price of Workplace Stress." *Harvard Gazette,* July 16, 2016.

[2] Bakker, Jorn, et.al. "What's your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data." (2011). *IEEE Computer Society*: 573-580

[3] Darrante et. al. "The Effect of Stress on Consumer Saving and Spending." *Journal of Marketing Research* (2016).

[4] "Understanding the Stress Response." Harvard Health Publications (2011). Web.

flexibility,[5] and prolonged increase in heart rate and blood pressure leads to increased risk for hypertension, and thus, increased risk for heart disease.[5] Prolonged elevated cortisol levels as a result of stress can lead to fat buildup in the body.[4] Cortisol increases appetite, causing an increase in caloric intake as well as the storage of unused nutrients as fat.[6] Stress can also lead to the development of poor coping mechanisms, including drinking, smoking, drug use, and overeating, all which affect overall health negatively.[7] Both the direct effects, such as increased heart rate, and coping behaviors individuals use to handle stress, such as poor eating and sleeping habits, can lead to illness and a general overall decline in health. Knowing what leads to stress and how to reduce or prevent it would have enormous benefits for the health of individuals and society as a whole.

Some physiological responses that have been previously used to attempt to predict stress include heart rate and galvanic skin response.[8] Data for these metrics can be collected through the use of different sensor mechanisms, for example, the Microsoft Band, which is the sensor used for this study. Currently, attempts to identify a person's stress level through sensors has proved challenging, partially due to the lack of situational context surrounding the data obtained through sensors.[9] Improved methods would provide more accurate prediction of stress and enable more effective intervention to mitigate the negative effects of stress on individuals.

Normative Analysis

While predicting stress using sensors alone had proven challenging, analyzing sensor data in addition to other data points should improve prediction accuracy. Not only will we examine data such as heart rate, exercise level, and skin conductance, but also user surveys on stress levels. Using the individual surveys in addition to the sensor data we can get a more accurate prediction of whether these factors contribute to an individual's stress level.

Increasing the accuracy of predicting stress is vital to being able to reduce the negative effects stress has on individuals. A wearable sensor could alert a user when they are likely to become stressed, and recommend methods to reduce stress levels. Doctors could do something similar by monitoring patients' heart rates and other factors in order to provide interventions before a patient's stress reaches an unhealthy level. Companies could also monitor average stress levels and respond to decrease them or provide access to healthy coping mechanisms when an employee is alerted to having high stress levels. This could prevent the cost accumulated due to stress-related workday loss. Being able to predict what will contribute to stress will help individuals be able to prevent and reduce stress levels to a healthier level and lead better lives.

---

[5] Reuel et. al., "Heart Rate and Blood Pressure: Any Possible Implications for Management of Hypertension?" HHS Public Access (2012).

[6] Herbert, Cohen et. al. "Stress and Illness." *Encyclopedia of Human Behavior*, Vol. 4 (1994).

[7] Jackson et. al. "The National Survey of American Life: a study of racial, ethnic and cultural influences on mental disorders and mental health." Int J Methods Psychiatry Res. (2004) 13(4):196-207.

[8] Liu et. al. "Listen to Your Heart: Stress Prediction Using Consumer Heart Rate Sensors." (2014).

[9] Bakker, Jorn, et.al. "What's your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data." (2011). *IEEE Computer Society*: 573-580

Stakeholders and Impact

All of us have a stake in being able to reduce stress. Stress can lead individuals to have poorer health overall, which leads to a lower overall quality of life.[10] Businesses have an interest in reducing the stress-levels of their workers in order to increase productivity and reduce loss due to worker absences caused by stress.Life-coaches, social workers, teachers, and personal trainers also have an interest in predicting client stress, in order to better improve the lives of their clients.

Providing individuals with the information to reduce and prevent stress will increases wellness, worker productivity, and overall well-being. If individuals and healthcare providers can predict when a person will become stressed, then interventions can be made in order to mitigate the effects of stress on a person's overall health.

Objectives and Metrics

In this research we will attempt to increase the prediction accuracy of whether or not a person will become stressed in the next 24 hours. Ideally, this would lead to individuals having a better understanding of what causes stress and being able to use this information to prevent and reduce the negative consequences of stress.

We will use mobile sensor data collected from volunteers, in particular, data surrounding heart rate, steps taken, skin conductance, and volunteer surveys on their stress level. We will build various models using this data, and use cross-validation to select the model with the highest accuracy.

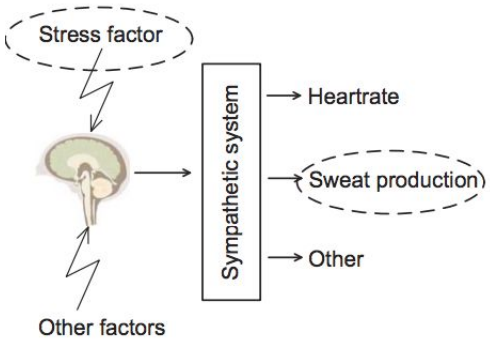
Literature Review

In 2011, researchers from Eindhoven University of Technology tried to detect human stress by measuring Galvanic Skin Response (GSR) data (reflecting sweat) measured by a watch-like device worn by subjects. The focus of the study was the identification of short term stress (acute stress), specifically the detection of a change of state, in which the study assumed a subject could only experience one of two states: *normal* or *stress*. The study noted however that using sweat as a response metric may have some issues not only due to the noise of the collected sensor data (such as how tight a subject wears the device), but also more importantly due to the fact that other factors other than stress (such as adaptation to a change in temperature) can cause the sympathetic system to produce sweat (see figure below).[11]

---

[10] Pazzanese, Christina "The High Price of Workplace Stress." *Harvard Gazette,* July 16, 2016.

[11] Bakker, Jorn, et.al. "What's your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data." (2011). *IEEE Computer Society*: 573-580



This data collection was conducted over the course of four weeks, with five human subjects wearing the watch-like device during working hours. Since the sampling rate was 4 Hz, and the typical working day is roughly 8 hours, the average length of the raw time series was 98721, in which formed a dataset of 72 time series. However, after preprocessing the data, 26 time series were excluded from the analysis due to either the GSR level showing very low

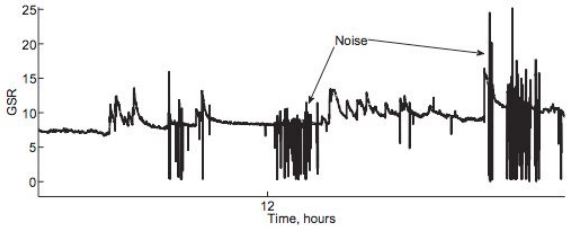


Fig. 5. The GSR signal contains two-sided local noise peaks that are probably caused by a physical disturbance of the contact between the skin and the sensors, e.g. if someone has a habit to touch from time to time the watch or the stress meter in this case.

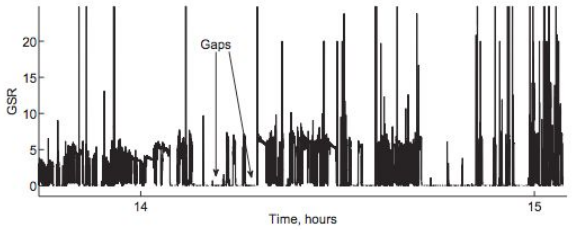


Fig. 6. When the fit between the skin and the sensors is not tight enough, the contact is continuously broken. A characteristic of this behavior is the high amount of gaps (ground value of sensor) in the signal.

variation or that the contact of the sensors was not adequate enough to yield reliable signals (which was detected by a filter and verified by visual inspection). For each of the remaining 56 time series state change points were annotated based on the visual inspection, which represented their ground truth. Overall the set of time series contained 368 change points with an average of around 6.5 change points per time series.[12]

[12] Bakker, Jorn, et.al. “What’s your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data.” (2011). *IEEE Computer Society*: 573-580

As mentioned before, before any model formulations, data was preprocessed through a three-step procedure. First, noise caused by contact loss or physical disturbance with the device was removed from GSR time series (see figure below). Local physical disturbances were filtered out by using a median (rather than a moving average) filter in order to preserve edges while filtering out noise.[13]

In the second step, filtered values were aggregated via a moving window of size 100. In the final step, data was discretized using Symbolic Aggregate Approximation (SAX) into a time series of values ranging from 1 to 5 (1 being completely relaxed and 5 being maximally aroused). The study warned that interpretation of these discretized time series should be local relative measures of arousal rather than absolute levels of arousal. The figure below highlights the complete preprocessing procedure.[14]

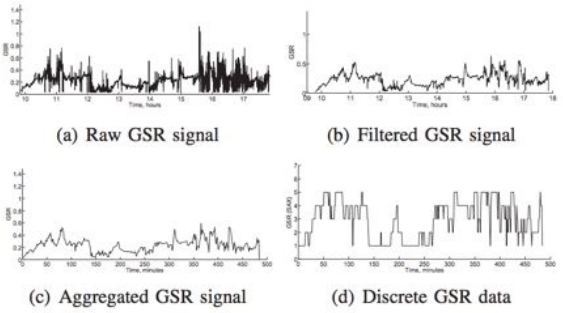


Fig. 12. An example of GSR signal in its original form and after each of the three individual steps in the data preprocessing: the raw GSR signal shown in (a) is filtered using a median filter (b), then the values are aggregated to the minute level (c), and finally they are discretised using SAX encoding (d) to be used as an input for a change detection technique.

In order to detect a change in state, the study used two approaches. In one approach, the study used is Adaptive Windowing (ADWIN), in which the length of a variable-length window of recently seen data points is maximized and is statistically consistent with the hypothesis that there has been no change in the mean signal value inside the window. Methodologically, given a sequence of signals, check whether there are statistically significant difference between the means of each possible split of sequence. If a statistically significant difference is found, the oldest portion of the data backwards from the detected point is dropped and the splitting procedure is repeated until there are no significant differences in any possible split of the sequence.[15]

In the other approach (*Fit)* monitors the mean of the input data and signals when it is statistically significantly different from zero (when the data deviates from the normal process behavior). Given historic preprocessed data, the objective is to fit a simple regression model. Based on the MSE for the incoming data, a statistical measure (the Mann Whitney U test in this study) determines whether there exists a significant change in prediction. Every time a new datapoint arrives, the data is split into two sets: a reference set that excludes the new point and

[13] Bakker, Jorn, et.al. “What’s your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data.” (2011). *IEEE Computer Society*: 573-580
[14] Bakker, Jorn, et.al. “What’s your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data.” (2011). *IEEE Computer Society*: 573-580
[15] Bakker, Jorn, et.al. “What’s your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data.” (2011). *IEEE Computer Society*: 573-580

the test set that also includes the new point. For each set, the model is trained by using Leave-One-Out. If there is an overall significant difference between the two sets, it is considered to be a change point and a cut is made.[16]

For each 56 time series, the study preprocessed the data, applied the two change detection methods, and compared the classifications of the changes detected by each method.The change points were evaluated by measuring the distance between the point identified by the detection algorithm and the closets actual change point within a predetermined boundary threshold. Measures of True Positive and False Discovery rate within a window of 5 minutes around the actual change point were recorded. False Discovery rate was used instead of False Positive rate due to the large amounts of True Negatives with respect to the True Positives.[17]

Results showed that neither methods were able to detect all change points. However, *Fit* detected more change points than *ADWIN*, but at the cost of more False Positives. *ADWIN* was better at positioning the change points. An example of change detection for one time series is shown below.[18]

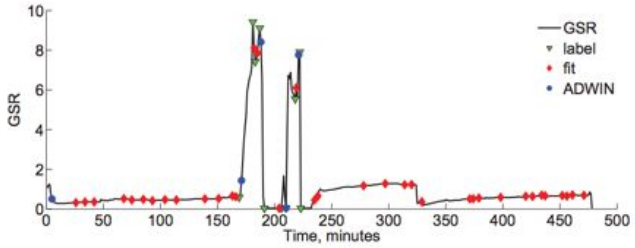


Fig. 15. One of the stress time series and the change points. Green triangles depict the ground truth, red diamonds depict the detection of the *fit*-method, and the blue circles depict the detection of *ADWIN*.

ADWIN did not detect small peaks and had a hard time detecting changes in cases where the signal slowly rose or fell, thus resulting in a lower True Positive rate compared to *Fit*. The study also verified their use of discretizing the data via SAX by showing that change detection became less accurate when used on non-discretized data. In the two figures below, one can see that *ADWIN* missed three change points.[19]

---

[16] Bakker, Jorn, et.al. “What’s your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data.” (2011). *IEEE Computer Society*: 573-580

[17] Bakker, Jorn, et.al. “What’s your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data.” (2011). *IEEE Computer Society*: 573-580

[18] Bakker, Jorn, et.al. “What’s your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data.” (2011). *IEEE Computer Society*: 573-580

[19] Bakker, Jorn, et.al. “What’s your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data.” (2011). *IEEE Computer Society*: 573-580

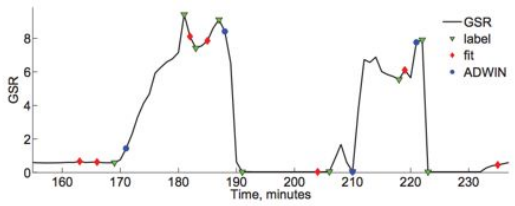


Fig. 16. Closeup of the time series in Figure 15. *ADWIN* clearly detects the high peaks, whereas the *fit* method is more sensitive to small local changes.

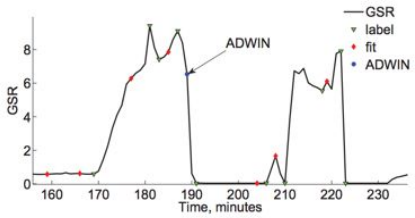


Fig. 18. Detection results of *ADWIN* and *fit* on the same time series as in Figure 16, but without SAX discretisation in preprocessing.

Overall, *Fit* was more sensitive to small local changes that *ADWIN* misses, but had a lot more False Positives than *ADWIN* (see tables below).[20]

TP AND FP RATES OF DETECTING THE CHANGE POINTS. THE MEAN μ VALUES ARE PERCENTAGES WITH RESPECT TO PERFECT DETECTION.

	$\mu(\frac{TP}{P})$	$\sigma(\frac{TP}{P})$	$\mu(\frac{FP}{TP+FP})$	$\sigma(\frac{FP}{TP+FP})$
Fit	0.66	0.16	1.66	0.16
ADWIN	0.08	0.01	1.01	0.1

TABLE III
THE DISTANCE BETWEEN THE TIME OF THE ACTUAL CHANGE (t_a) AND THE TIME OF THE DETECTION (t_d).

	$\mu(\lvert t_a - t_d \rvert)$	$\sigma(\lvert t_a - t_d \rvert)$
Fit	2.8	0.54
ADWIN	2.5	1.2

The study also warned that interpreting GSR data was ambiguous. Even with “ideal” noise-free GSR data, a peak in GSR data might not correspond to stress (such as when a person is exercising). In order to further improve on the ideas of this research, we intend to use other variables, such as survey data of whether the participant is exercising and their tiredness level, along with the Microsoft Band heart rate, pedometer, and steps data.

The idea of a growing desire for oneself to understand reasoning behind specific behaviors and biological events was also studied by Tom Fawcett of *Silicon Valley Data Science* [21]. This idea of “Quantified Self” data refers to all data which is measured by various self monitoring apps or tools to allow for individuals to record information about oneself. Rather than simply observe the data or use it to predict certain events like many currently do, Fawcett discusses ways in which individuals can find knowledge about reasoning behind certain events based on data pertaining to their health. Instead of simply trying to predict when or if an event such as a migraine will occur, this data can be used to observe factors that may be related to this occurrence. Other such events Fawcett believes can be integrated into this sort of analysis include anxiety attacks, nights of poor sleep, periods of high energy, periods of low motivation, and periods of high irritability.

Fawcett uses data compiled by users of such tracking devices to collect approximately six months worth of data on various data types. Some of the variables analyzed are self-reported by the user, meaning they give themselves a score, such as a health score, energy level, mood ratings, happiness and life satisfaction. Other variables were measured directly from a health app such as stress level, alertness, sleep duration and efficiency, calorie

[20] Bakker, Jorn, et.al. “What’s your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data.” (2011). *IEEE Computer Society*: 573-580

[21] Fawcett, Tom. "Mining the Quantified Self: Personal Knowledge Discovery as a Challenge for Data Science." Big Data 3.4 (2015): 249-66. Web.

intake, time spent active, and number of steps taken. Any variables that required it, were then averaged together to create a single observation for each day over the six month testing period. This data was collected as a time series, but in order to perform the desired analysis Fawcett performed discretization techniques in order to transform the data into discrete variables. A similar SAX technique that was discussed in Bakker's paper was implemented to break the time series into equally sized windows of time. These blocks of time are referred to as 'episodes' which were then analyzed and classified based on the observed data within each episode.

As Fawcett was trying to derive knowledge of why events happen rather than prediction, he implemented an unsupervised technique to cluster the episodes of data. Using a 'k-mean algorithm,' the days were clustered into 5 different centroids based on what a person experienced or felt each day. These clusters were classified as inactive days which consisted of low activity, low energy, or low mood, high activity days consisting of high activity, energy, and mood, moderate activity days, depressive or high stress days consisting of low sleep efficiency, a large amount of sleep and low alertness, and finally high stress, high intensity days, but not a lot of activity. It was

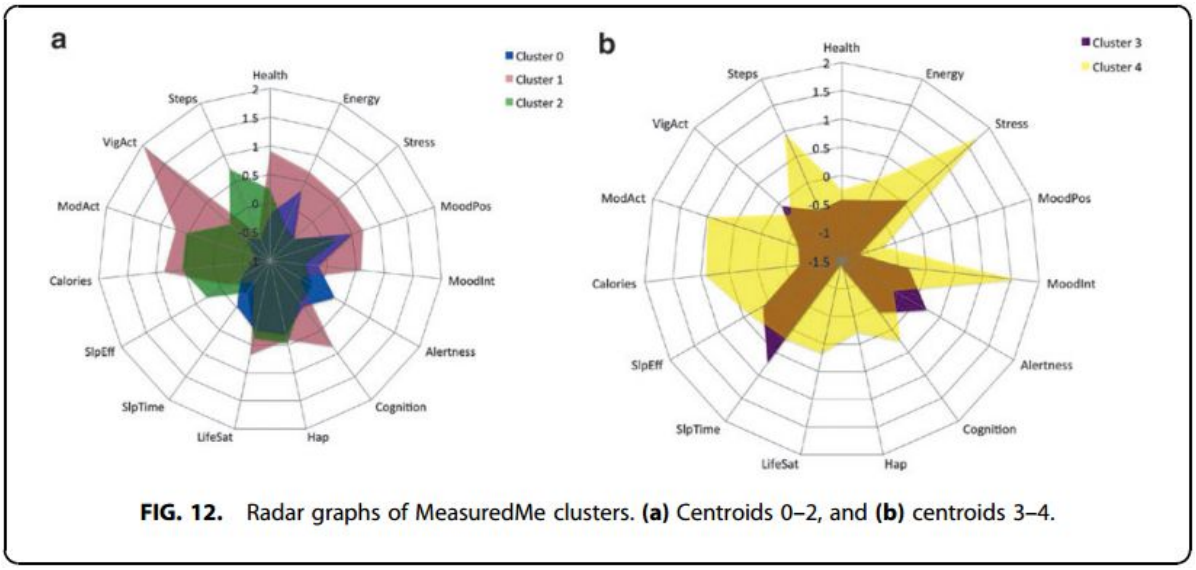


FIG. 12. Radar graphs of MeasuredMe clusters. **(a)** Centroids 0–2, and **(b)** centroids 3–4.

discovered that the clusters with the highest concentration of days were those described as inactive days and those classified as moderately active days. While this analysis allows for users to further understand how their days are being spent, due to the discretization of the data, it is impossible to fully examine how a previous day's classifications affect future days.

Hypotheses

Our hypothesis is that fitness sensor data is predictive of a user's future stress level. In order to test our hypothesis, we define our null hypothesis as fitness sensor data not being predictive of a user's future stress levels. Our alternative hypothesis is that the fitness sensor device data is predictive or the user's future stress level. We will determine that our model is predictive if it statistically significantly beats the NAIVE model; the NAIVE model predicts the stress level of all users as the same stress level, based on the most frequent stress level response.

Data

The data was collected over a period from August 25th, 2016 to November 8th, 2016 through the Sensus application designed by the UVA Predictive Technology Laboratory assisted by Microsoft Bands. Members of the Systems 6018 class volunteered to participate and a total of 19 devices were used to collect data from an unknown subset of human participants. The app and bands collected close to a gigabyte of data over this time period, however we immediately removed some data such as cell tower proximity and timestamps of phone calls or text messages sent or received. After removing this data we focused largely on the data connected by the Microsoft Bands, as well as some readings taken through a user's phone. Among this data were the results of three surveys that each participant was prompted with over the collection period. Our first step was to clean these surveys as these contained our response variable.

The three surveys collected over the time of the collection were "On a scale of 1-5 how stressed are you?", "On a scale of 1-5 how tired are you?", and "Are you exercising?". The first two were prompted at random times throughout the time while the latter was collected when the Microsoft Band produced a heart rate reading greater than 100 bpm. Our focus was on user's stress, therefore after loading in the JSON files within the ScriptDatum folder which contained the results of all three surveys, our first step was to remove the surveys related to tiredness and exercise. This left us with 34 observations to look at. To further clean the dataset, all columns except the user's response, the timestamp of the user's response, the device ID, and the latitude and longitude of the user were removed. After analyzing the final dataset, it was observed that the surveys that fit our criteria were only collected between November 2nd and November 8th, 2016. The timestamp and the device ID were kept in order for us to associate other recorded data with a user's response to the survey. Finally, Professor Matthew Gerber had two devices recording data, one Android and one IOS, therefore we needed to remove one of these devices. Due to the differences in the available data collected between Androids and IOS phones, we decided to keep all of Professor Gerber's Android data and removed the IOS observations. We would go on to perform this removal of his IOS data in all following datasets as well.

The next dataset of concern was produced from the JSON files within the Microsoft Band Contact Datum folder. This data displayed whether the user's Microsoft Band was within contact of the wearer's skin. The data resulted in three different states: having contact with the user, not having contact with the user, and maybe having contact with the user. This data was taken approximately once every 1.5 seconds and we removed all observations where the state was not definitely having contact with the wearer. We felt this would allow us to produce the most accurate results as this data would make certain that the bands were reading accurately and transmitted to the device. After removing unused variables, the final dataset consisted of the status as previously discussed, the timestamp of the reading, and the associated device ID. This dataset will be used to find accurate readings of other variables related to the Microsoft Band.

Related to the Microsoft Band, the next variables we looked at were distance traveled, heart rate, steps ascended, total steps and galvanic skin response (GSR). Each of these variables were read in similar to the previously discussed data sets as JSON files. All five datasets were then cleaned in similar fashion. All variables were removed aside from the timestamp of the reading, the device ID, and the associated reading's result. Since all surveys of interest were recorded in November, we removed all data that was not recorded during this time frame. The next step was to associate these reading with a contact reading from the previously discussed dataset. This ensured that every reading was as accurate as possible by the Microsoft Band's standards. However, as we will discuss in the biases, this still was not 100% accurate for a reading.

In order to investigate more predictors, we also did similar cleaning for several variables related to the user's device; the battery life and the volume level. The battery life was measured as a percentage of the phone's overall life while the volume was measured in decibels. Since these readings were recorded directly through the user's phone it was not necessary to match them with a contact reading and therefore any reading that took place within the month of November was kept.

Variable	Description
Response	Survey response to 'Rate your stress.' Answer could range discretely from 1-5. Actual responses ranged from 1-4.
Average_Steps_Ascended	Average steps ascended over the 24 hours prior to the user's survey response.
Average_Heart_Rate	Average heart rate over the 24 hours prior to the user's survey response.
Average_Total_Steps	Average total steps over the 24 hours prior to the user's survey response.
Average_GSR	Average GSR over the 24 hours prior to the user's survey response.
Average_Distance_Traveled	Average distance traveled over the 24 hours prior to the user's survey response.
Average_Battery_Life	Average battery life over the 24 hours prior to the user's survey response.
Classroom	Did the user come within 50 meters of the SYS 6018 classroom within the 6 hours prior to the survey being responded to. (1 if yes, 0 if no)
Average_Sound	Average sound level over the 24 hours prior to the user's survey response.

Table 1: Variables and descriptions

Finally, the app recorded a timestamp whenever the user was within 50 meters of a previously specified point of interest. This point was determined to be the classroom where lectures for the student's Systems 6018 class are held. Similarly, the readings were cut down to only include those in November and contained the timestamp and the device id of the user. These recordings were most likely observed due to two distinct occurrences. Either the user was attending their Systems 6018 Data Mining lecture or they were attending their Data Science 6002 Ethics of Data Science class which took place within 50 meters of the former's classroom. For our purposes we did not find it necessary to distinguish between the reasoning for the occurrence of the reading.

In order to associate these readings with a user's survey response, we needed to distinguish a window around each survey. After analyzing the possibility of 1, 2, 3, or 24 hour windows prior to the survey response, we determined a 24 hour window was necessary. Anything less would not provide sufficient observations of the predictor variables due to the removal of any data point not matching our previously specified criteria. However for the data for being within the proximity of the classroom was cut down to a window of 6 hours prior to the survey. For the variables of heart rate, steps ascended, total steps, distance traveled and GSR, the average value of all reading in the 24 hours prior to the survey were averaged to produce an average rate. A similar calculation was performed for battery life and phone volume. For the proximity to the classroom variable, a binary true or false variable was created to state if the user was within the range during the 6 hours preceding the survey.

Due to issues with recording the data that will be discussed in the biases, some surveys did not contain values for several predictors, either entirely or for specific variables. In order to create a more complete dataset, random forests were used to impute the missing values. This method did still contain several concerns as it was possible for the data which we trained our random forest on to not contain certain variables due to missingness and therefore be unable to predict accurately. Therefore to ensure we were as accurate as possible, we imputed the missing values by conducting 100 trials of the random forest with m= $\sqrt{p}$ and the resulting missing values from each trial were averaged together.

Biases

We acknowledge that our data contains many biases. After significant data cleaning and processing, we were only left with 34 observations of stress survey responses from individuals wearing the sensor bands. This was due to the fact that few individuals participated in the study to begin with, even fewer were wearing sensor bands, and even fewer still regularly answered the stress surveys. This creates a reference problem since 33% of the survey responses came from one individual and the results are slanted toward that individual's characteristics. Having so few observations will cause the variance of our model to be high. In addition, for the 34 observations that we did have, we did not have complete data for many of our predictors of interest. Out of the 34 observations, only 9 were complete. Since creating a model with only 9 observations would have extremely high variance, we decided to impute the missing values using a random forest imputation R package. While this is not ideal, it is still preferable to creating a model with only 9 observations.

In addition to the few complete observations in the data, the data we were able to collect was a cause for concern. Many of the values seem to suggest that the bands are inaccurate in their readings. For example, despite the fact that we filtered the band data to only observations with a contact reading of "2", defined as positive contact with the band, there were still heart rate readings that were too low to be accurate. It is impossible to have a heart rate of 0 and unlikely that someone's actual heart rate would be less than 40. While technically it is possible for someone's resting heart rate to be 40, according to the Mayo Clinic, this is only possible for a well-trained athlete, which we know does not apply to our sample since we personally know the few individuals who were wearing the bands and none are well-trained athletes.[21] Despite this, we had one observation of 0 and and many for less than 40 for heart rate. This would obviously skew our data when computing the average heart rate in a 24 hour period, however, we decided to leave these observations in the model due to the fact that we had such little data to begin with and we were concerned that removing these observations would only increase our model's variance even further. There were also observations for GSR of 0 even though we filtered out observations where the user was not in contact with the band. For these GSR observations, we decided to treat values of 0 as missing values and impute them using the random forest imputation R package since there were so many of them.

Another issue with the data was that despite the fact that the stress survey response options were "1,2,3,4,5", with 1 being not stressed and being very stressed, our data did not contain any responses of 5. Since there were no responses of 5 in our training data set, we are unable to predict when someone will respond "5" to the stress survey questions. This illustrates a significant flaw in the model, the fact that we are unable to predict one level in the 5-level stress survey. Having more observations, conducted over a longer period of time and with more participants, would likely cause our data to include individuals who reported a stress level of "5." This would also likely increase the accuracy of the model as our data set increased both in terms of number of observations and scope of stress survey responses. In addition, the survey responses on a 1-5 scale may mean different things to different individuals, so individuals with similar levels of stress may report different numbers on the 1-5 scale. These inconsistencies could lead to greater inaccuracy of the model.

Methods

After cleaning and imputing our data, we created several models: a NAIVE model, three random forest models, and a linear discriminant analysis model. As mentioned before, the NAIVE model predicted the stress level of all subjects as a level of 3 since it was the most frequent response.

---

[21] Laskowski, Edward. "What's a Normal Resting Heart Rate? (2015) http://www.mayoclinic.org/healthy-lifestyle/fitness/expert-answers/heart-rate/faq-20057979

Response	**1**	**2**	**3**	**4**	**5**
Frequency	2	8	20	4	0

Table 2: Response frequency to stress survey

In order to compare accuracy of our NAIVE model with the other models we built (random forest and LDA), we randomly selected 25% of our data as a testing data set and predicted the accuracy of our NAIVE model on the testing data set. We repeated this process 100 times and took the average accuracy rate. The average accuracy rate of our NAIVE model was 0.594. This did not differ much from the raw accuracy rate of our NAIVE model, which was 0.588 (20/34), based on the full data set.

The next model we built was a linear discriminant analysis model. This model created the different responses to the stress survey (1, 2, 3, or 4) as individual classes that are grouped together by different characteristics in the predictor variables, essentially aiming to maximize the component axes for optimal class separation. First, variable selection was performed using a stepwise method. This process calculated the variables that resulted in the best model accuracy, where the final result was the model that used the variable for **Average_Steps_Ascended** and **Average_GSR** as predictors. This is novel compared to prior work in that other studies have mostly focused on only GSR or heart rate to predict stress, where we are using variable selection to find other variables that may be not be as obvious. Assumptions for an LDA model include that the observations are random and that the variables are normally distributed. These assumptions are satisfied by our data.

Next, random forest models were built by examining various response variable. We varied *m (*the number of sub predictors to examine) in our random forest model and tested accuracy ratings for $m=\sqrt{p}$, $m=p$, and $m=1/3p$. We created 250 different random forests, with 500 different trees for each random forest for each m value. Within each 250 different random forests, we used a random observation as our testing set, and the other 33 observations as our training set since only two observations had a response value of 1, and we wanted to guarantee that at least one of these responses is in the training set. After finding optimal *m* we built three different random forest models: the first model used all predictors in table 1, the second model only used physiological data as predictors, namely **Average_Steps_Ascended, Average_Heart_Rate, Average_Total_Steps, Average_GSR,** and **Average_Distance_Traveled**, and the third and final model examined variables **Average_Steps_Ascended** and **Average_GSR**, which were determined by the stepwise subset selection when we ran our LDA model.

This method is novel compared to prior related work since based on our literature review, we did not come across a study that used random forests to predict stress while looking at the variables that we looked at, in addition to comparing this result to a NAIVE model. Our assumption of using $m=\sqrt{p}$ as the optimal *m* was met, as described in the evaluation and results section.

Evaluation and Results

In order to evaluate the performance of the LDA model, leave-one-out cross validation was used to determine the prediction accuracy of the model. We found that the LDA model using the variables **Average_Steps_Ascended** and **Average_GSR** correctly classified the stress responses 64.7% of the time, but had a 95% accuracy confidence interval of (0.382, 0.911). Since the accuracy of the NAIVE model is within the CI of the LDA model, we conclude that the LDA model does not statistically significantly predict stress level better than the NAIVE model.

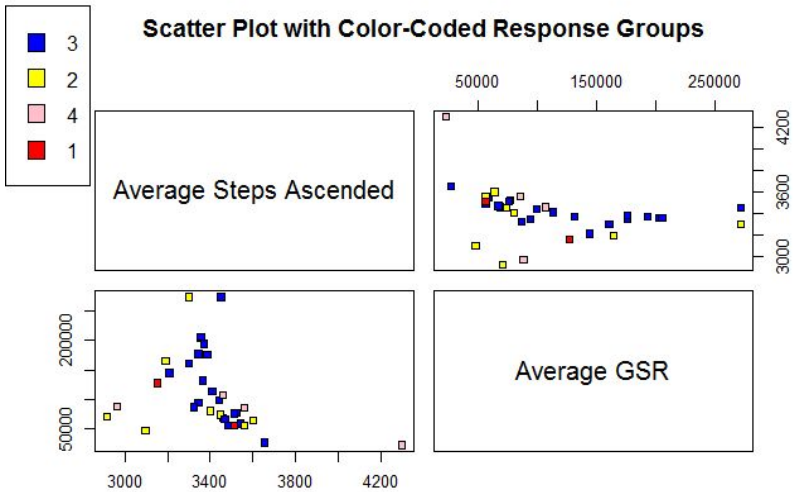
A confusion matrix was created to evaluate the specificity and sensitivity of the model.

		Predicted			
		1	**2**	**3**	**4**
Actual	**1**	0.0	0.0	1.0	0.0
	2	0.0	0.25	0.75	0.0
	3	0.0	0.0	1.0	0.0
	4	0.0	0.25	0.75	0.0

Table 3: Confusion Matrix for LDA Model

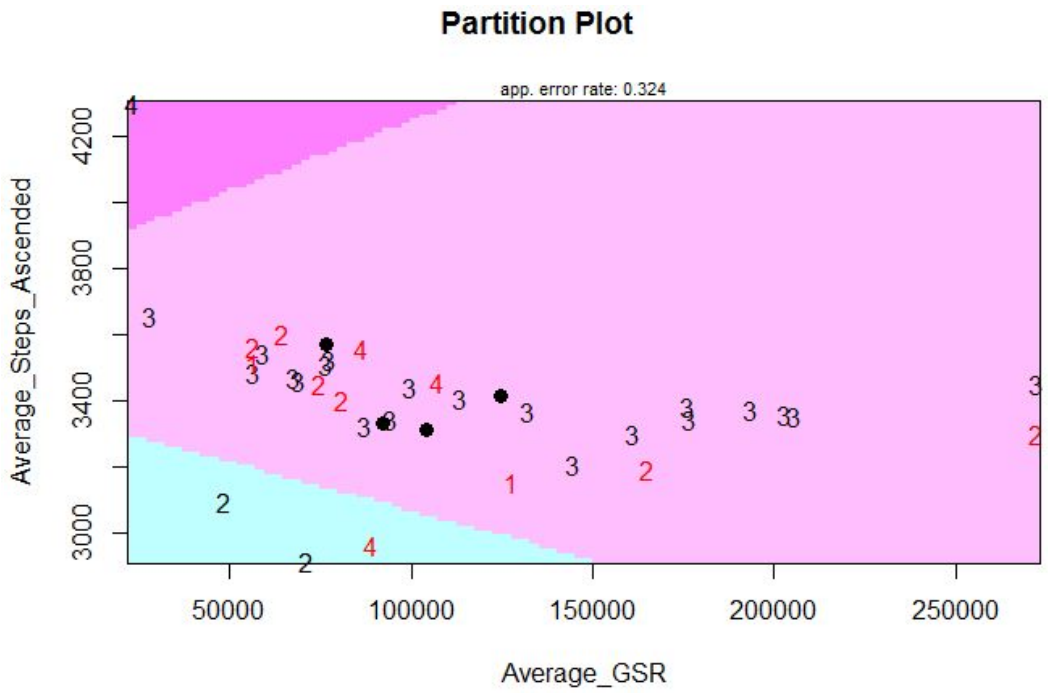
As is seen in table 3, the model is prone to over-classifying the stress response of ‘3’. It fails to predict both observations of ‘1’ and ‘4’. The model predicts a class of ‘3’ 100% of the time for actual observations of ‘1’ and predicts a ‘3’ 75% of the time for actual responses of ‘2’ and ‘4’. This is due to ‘3’ being the most frequent response in our data. There were so few of the other values that the model was trained to predict mainly for the ‘3’ class.

To further evaluate the model, visualizations were created to observe the class separation established by the model.



Plot 1: Scatter plot of actual survey responses for Average Steps Ascended and Average GSR

Plot 1 shows the different response classes grouped by color. It shows each of the variables used in the model plotted against each other. As is in the plot, there is no clear separation of the classes. This makes it very difficult to classify each response based on the characteristics, since there is not enough distinction to fully group the responses. It is also seen that '3' is the most frequent response, which explains the model's tendency to classify responses as '3' for the majority of predictions. In addition, we created a partition plot to evaluate the class separation.



Plot 2: Partition plot using an LDA model for Average Steps Ascended and Average GSR

Plot 2 shows the areas allotted to each class by the model. The black numbers are classes that are correctly in the designated area. The red numbers depict observations that are in the incorrect area created by the model. The most prominent area is the '3' area which is light pink. You can see the few correct classifications of '2' performed by the model are in the light blue area. Many of the responses were clustered in the category designated for a response of '3', which further shows that the data was difficult to group by class, diminishing the prediction accuracy of the model.

For our random forest models, we found that $m=\sqrt{p}$ had the highest average accuracy of 0.493, and that m=p had the second highest accuracy of 0.48 and m=1/3p had the lowest average accuracy of 0.439. From this point forward, we used $m=\sqrt{p}$ for our random forest models, on the assumption that $m=\sqrt{p}$ would have the highest accuracy on all three random forest models.

The first random forest that used all predictors had average mean test accuracy of 0.449. The second random forest that used physiological data as predictors had a mean accuracy of 0.493. The final random forest model had a mean accuracy of 0.449.

Because our LDA and random forest models did not out-perform the NAIVE model, we fail to reject our null hypothesis that sensor data can predict an individual's stress level. However, our work shows that further research in this study with more accurate data could potentially lead to more accurate results. With more accurate results, it could be possible to not only predict an individual's stress level but also intervene in order to prevent or reduce stress levels to help individuals lead a longer and healthier life.

Being able to predict a person's future stress level is challenging, and likely involves other factors that were not available to us for this study, including diet and workload. A future study that looked at these and other factors in addition to the sensor data in this study may result in a more accurate model.

There are many improvements that can be made in quality and accuracy of sensor related data. This was not just something that we observed but that was also observed in many of the other studies we reviewed. In addition, it is challenging to consistently collect data from volunteers. It was difficult to get volunteers to wear the band to begin with and even those individuals who volunteered did not all wear the band and answer the survey questions consistently.

Future Research

While there were many limitations faced with this research due to the nature of the experiment and therefore the resulting data, our results showed some promise. While the prediction accuracy of the models was lower than hoped for, there appeared to be some finding which could be further explored. Given a more structured experiment which recorded readings on a more constant and confident basis, a smaller window than 24 hours could have been implemented. We would then have been able to cross-validate to determine the optimal window for stress level prediction. There are also many other variables not investigated which could influence an individual's possibility of becoming stressed in the near future. Some such variables could include sleep time, age, and other demographic variables that could be influential to someone's stress level. Any future research may also want to attempt to collect data over a longer time period in order to collect what will most likely be more accurate data. ú

References

Pazzanese, Christina “The High Price of Workplace Stress.” *Harvard Gazette,* July 16, 2016.

Bakker, Jorn, et.al. “What’s your Current Stress Level? Detection of Stress Patterns from GSR Sensor Data.” (2011). *IEEE Computer Society*: 573-580

Darrante et. al. “The Effect of Stress on Consumer Saving and Spending.” *Journal of Marketing Research* (2016).

“Understanding the Stress Response.” Harvard Health Publications (2011). Web.

Reuel et. al., “Heart Rate and Blood Pressure: Any Possible Implications for Management of Hypertension?” HHS Public Access (2012).

Herbert, Cohen et. al. “Stress and Illness.” *Encyclopedia of Human Behavior*, Vol. 4 (1994).

Jackson et. al. “The National Survey of American Life: a study of racial, ethnic and cultural influences on mental disorders and mental health.” Int J Methods Psychiatry Res. (2004) 13(4):196-207.

Liu et. al. “Listen to Your Heart: Stress Prediction Using Consumer Heart Rate Sensors.” (2014).

Fawcett, Tom. "Mining the Quantified Self: Personal Knowledge Discovery as a Challenge for Data Science." *Big Data* 3.4 (2015): 249-66. Web.

Paoli, Pascal, et. al. “Third European survey on working conditions 2000.” *European Foundation for the Improvement of Living and Working Conditions* (2001).

J. Lin, E. J. Keogh, L. Wei, and S. Lonardi. “Experiencing SAX: a novel symbolic representation of time series.” Data Min. Knowl. Discov., 15(2):107–144, 2007

Laskowski, Edward. “What’s a Normal Resting Heart Rate?” (2015) http://www.mayoclinic.org/healthy-lifestyle/fitness/expert-answers/heart-rate/faq-2005797