# Listen to Your Heart: Stress Prediction Using Consumer Heart Rate Sensors

David Liu, Mark Ulrich
{davidcyl, mark.ulrich}@cs.stanford.edu
Final Project, Stanford CS 229: Machine Learning, Autumn 2013-2014

**Abstract**—Galvanic skin response (GSR) is commonly used to measure short-term emotional and cognitive stress, but measuring GSR requires obtrusive equipment. We present a model to predict stress levels solely from electrocardiogram (ECG) data, which can be measured with wearable consumer-grade heart monitors. Our model incorporates time and frequency domain features of heart rate variability, and the spectral power components of the ECG. We make predictions on a sliding window of ECG signal. We apply a linear model to predict stress as a continuous quantity, and our prediction is correlated with the actual GSR with $R^2 = 0.873 \pm 0.035$. We also present a model for classifying each window as a binary "stress" or "rest" periods. The best performance was achieved with a linear SVM, with an $F_1$ score of 0.98 on the most distinct samples (highest and lowest 20 percentile GSR levels), and $F_1 = 0.85$ over all samples.

## 1 Background

Society and science alike agree on the deleterious effects of stress, which we will here define as the short-term activation of the sympathetic nervous system caused by cognitive stressors (as opposed to physical activity). Chronic emotional and mental stress has been linked to a range of health problems [1], [2]. People often don't notice triggers that cause them to become stressed. Stress is easy to identify using galvanic skin response (GSR), which measures sweating of the hands. However, these sensors interfere with daily activity. Meanwhile, unobtrusive, wearable, off-the-shelf heart rate sensors and electrocardiogram (ECG) devices are becoming increasingly affordable. We seek to use this sensor data to determine when the wearer is experiencing stress.

A simplistic heart rate analysis usually takes a long time to detect a stressor, and it is difficult to determine if an increase in heart rate was caused simply by physical activity (such as standing up) or cognitive/emotional stress. On the other hand, GSR is particularly useful in the immediate identification of short-term stress events, because about 1-2 seconds after a startle event, an individual's palms will sweat and GSR will spike [3]. The goal of our project is to achieve the accuracy of GSR-based stress detection with only ECG signals.

## 2 Data Set and Features

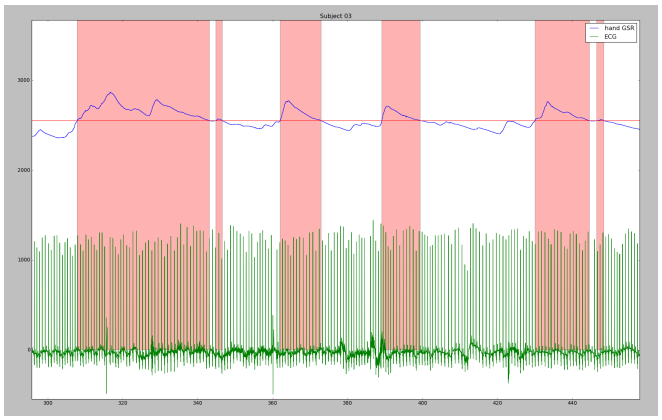



**Figure 2.1:** An example of GSR data (top) and ECG data (bottom). Shaded regions correspond to GSR values greater than 60% of all other GSR values in the record, an indicator that the subject was possibly experiencing stress.

We use data from a study by Healey and Picard [4], available from PhysioNet [5]. In this data set, GSR (on the skin of the hand) and ECG signals were continuously collected while drivers experienced stress-inducing events (busy streets, red lights, highways) as well as a rest state (parked in a garage). The signals were recorded at a rate of 31 Hz for the GSR and 496 Hz for the ECG. Healey's original study also included a variety of other features, such as respiration, electromyogram in the shoulder (muscle tension), and foot GSR. Hand GSR is the most commonly used indicator of stress, so we use that as the ground truth for subject stress.

This data is relatively noisy. In about 10% of the records, the data collection had clearly gone awry (detached leads, wildly noisy/jittery signals), so we manually examine the signal graphs and exclude those portions of the records. Each record is windowed using a fixed-size sliding window of length $W$ seconds (using cross-validation, we later determined that $W = 24$ was the

best option). From the ECG signal in the window, we extract heart rate variability features as well as spectral features of the raw waveform itself.

### 2.1 Heart rate variability features

Heart rate variability (HRV) features can be extracted from the timing of heartbeats alone. These features therefore only capture a limited subset of the information in the ECG signal, but they can be collected with only a simple heart-rate sensor, which is less expensive than a full ECG device.

We first use an automatic QRS wave annotation tool (WQRS) on the ECG data to identify morphological features in the ECG, such as the R peak of each heartbeat. The RR intervals (lengths of intervals between heartbeats) are then used to extract heart rate variability (HRV) features.

We use tools available through the PhysioNet WFDB package and HRV features that are commonly used for ECG analysis [6]. The time-domain features of HRV consist of the following characteristics of RR intervals:

- NN/RR: The fraction of heartbeats that are considered "normal" heartbeat lengths
- AVNN, SDNN, SDANN: Average and standard deviation of RR intervals
- SDNNIDX: Mean of the standard deviations of RR intervals in all 5-minute segments
- RMSSD: RMS difference between adjacent RR intervals
- PNN: Percentage of adjacent RR intervals that differ by more than 50 ms

To compute frequency-domain features of HRV we apply a Lomb periodogram, a variant of the Fourier transform designed for time series sampled at uneven intervals (such as heartbeats). The following features are extracted from the periodogram:

- TOTPWR: Total spectral power of all RR intervals up to 0.04 Hz
- ULF, VLF, LF, HF: Total spectral power of all RR intervals in the bands 0-0.003 Hz, 0.003-0.04 Hz, 0.04-0.15 Hz, 0.15-0.4 Hz
- LF/HF: Ratio of power in the LF to HF bands

### 2.2 ECG-based features

The raw ECG data contains the full spectrum of the heart's electrical activity, giving data that is richer than the timing of the heartbeats. We compute the Fourier Transform and take the logarithm of summed total power in 10Hz bands, from 0 to 200 Hz. This follows previous work by Chou et al. who showed that these ECG frequency bands were discriminative for detecting abnormal heart events [7].

## 3 Model

### 3.1 Continuous Stress Prediction

Our first approach was to apply linear regression directly to the features and attempt to predict a continuous stress value from the ECG features. We tried ECG spectra alone, as well as a combination of ECG and HRV features as predictors in the linear model. We experimented with different fitting procedures, including ordinary least squares, ridge regression, and LASSO regularization.

### 3.2 Stress Detection with Binary Classification

We next formulated the problem as a binary classification problem. Windows are labeled as "stress" or "rest" windows. To produce ground-truth labels, we use the relative value of the GSR signal: when GSR is in the top or bottom *cutoff* percentile in the record, the window is considered "stress". A *cutoff* of 0.5 corresponds to labeling all windows with above-median GSR as "stress" and all below-median GSR windows as "rest". When *cutoff* is less than 0.5, we train on and classify only the subset of data where the GSR is most extreme portion of the record: this corresponds to estimating stress only for the most clearly stressed or relaxed time periods.

We applied many techniques to this classification problem. The first was a Gaussian Naive Bayes classifier, which fits a Gaussian distribution to each feature independently of the others. That is, for the classes $y = +1$ (stress) and $y = -1$ (rest), we modeled the class distribution of each feature $x_i$ independently as

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\pi\sigma_y^2}\right)$$

We then applied an SVM to the binary classification problem and tried many kernels (rbf, polynomial with degrees 2-4, sigmoid, and linear). To determine the best regularization parameter and kernel and evaluate performance, 15-fold cross-validation was used.

We also tried $k$-nearest neighbor classification, in which each window is assigned based on a vote of its $k$ nearest neighbors in the feature space, and a random forest classifier.

### 3.3 Varying Prediction Windows

We hypothesized that heart-based predictions might lag the GSR because GSR exhibits a faster biological response. Thus, we tried using a shorter subportion of the window to assign the ground truth label of the window. Figure 3.1 shows the results of varying this label window length. Decreasing the window length generally decreased our $F_1$ score, so we settled on using the same window for input features and labels.

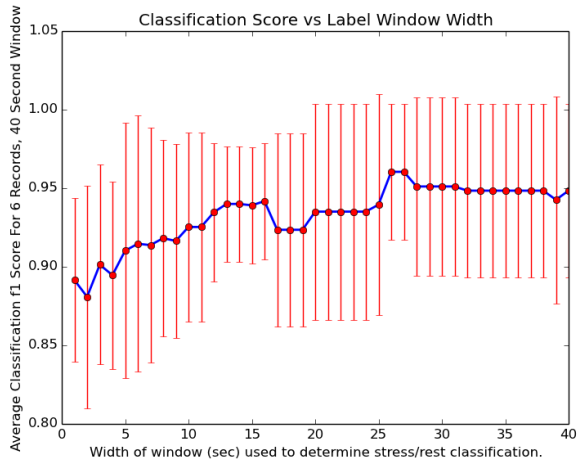



**Figure 3.1:** Performance under changing label window widths

We also tried a few more variations on the features and windows:

- We used previous and future ECG windows to predict the GSR in a short two-second window.
- To capture changes over the duration of the window, we extracted the HRV/ECG features separately from each third of the window, resulting in three times as many features.

Neither of these changes improved classification accuracy.

# 4 Evaluation and Results

## 4.1 Continuous Model Results

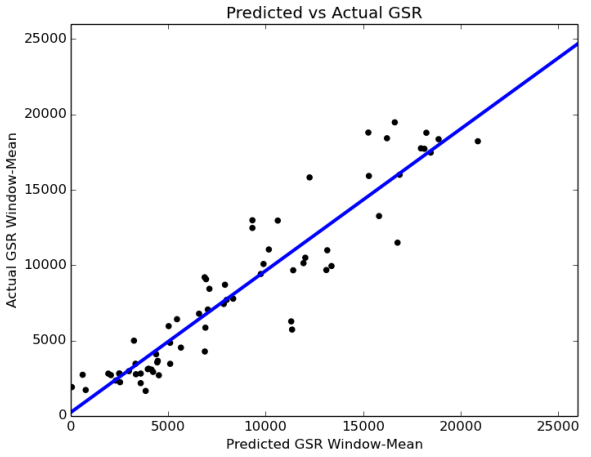



**Figure 4.1:** Scatter plot of ECG-predicted stress levels versus actual GSR, and the least-squares regression line. $R^2 = 0.87$. The regression line is $y = 0.9401x + 238.7$.

The linear regression yielded stress predictions that were reasonably correlated with the observed GSR. The model was evaluated with 15-fold cross-validation, and examining the correlation between the stress prediction ($x$-axis) and the actual GSR. A perfect prediction would have $R^2 = 1$ and all points on this plot lying on the line $y = x$. When predicting stress levels with only the ECG spectral features, we achieved $R^2 = 0.8474 \pm 0.05$. With both ECG and HRV features, we achieved $R^2 = 0.873 \pm 0.035$ and a line that is fairly close to unity (Figure 4.1).

## 4.2 Binary Classification Results

We measured the performance of our binary classifier using the $F_1$ score, which is defined as the harmonic mean of the precision and recall. That is,

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

| Classifier | Features | F1 Accuracy |
|---|---|---|
| SVM-rbf | hrv | 0.6071 |
| SVM-linear | hrv | 0.7268 |
| GaussianNB | hrv | 0.7855 |
| GaussianNB | ecg | 0.8447 |
| GaussianNB | ecg,hrv | 0.8815 |
| SVM-linear | ecg | 0.9787 |
| SVM-linear | ecg,hrv | 0.9836 |

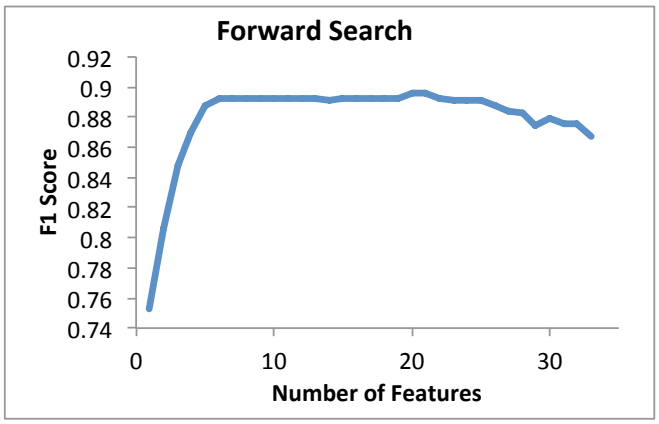
**Figure 4.2:** Selected combinations of classifier and feature set Classification with *cutoff* = 0.2. "ecg" corresponds to the 20 bands of ECG frequency content, and "hrv" corresponds to the 13 time and frequency domain HRV features.

We tried a range of different classifiers and feature sets. Figure 4.2 shows some selected combinations. We found that a linear SVM (i.e. logistic regression) on both ECG frequency features and HRV features performed best, outperforming other model choices slightly.

We also tried a $k$-means classifier and a random forest classifier, neither of which achieved competitive accuracy.

In Figure 4.3, we examine classification accuracy for varying *cutoff* GSR levels. In this graph, we see that classification accuracy is highest for the top and bottom 20 percent of GSR values, and classification accuracy drops off as we approach increasingly ambiguous GSR samples. This is reasonable, because the windows that have middling GSR values may not be as distinctively distinguishable as "stress" or "rest". When *cutoff* is extremely small, classification accuracy also suffers, though, because there are not enough training examples. When classifying all windows, the $F_1$ score is 0.84.

We varied the width of windows (Figure 4.4) and found that a length of 24 seconds is optimal. Extremely long windows performed increasingly poorly, likely because they often contained a mix of both stress and rest states.




| Feature | Score |
|---|---|
| ecg_freq_0_10 | 0.75277 |
| ecg_freq_150_160 | 0.80611 |
| ecg_freq_140_150 | 0.8475 |
| ecg_freq_10_20 | 0.86944 |
| ecg_freq_180_190 | 0.8875 |
| hrv_pnn | 0.89194 |
| **Additional features:** hrv_hf, hrv_lf, hrv_nn, hrv_sdann, hrv_totpwr, hrv_ulf, hrv_vlf, ecg_freq_160_170, ecg_freq_80_90, ecg_freq_50_60, ecg_freq_130_140, hrv_rmssd, hrv_avnn, ecg_freq_90_100, hrv_sdnn, ecg_freq_40_50, ecg_freq_120_130, ecg_freq_60_70, hrv_sdnnindx, ecg_freq_170_180, ecg_freq_20_30, ecg_freq_110_120, ecg_freq_190_200, ecg_freq_70_80, hrv_lfhf, ecg_freq_100_110, ecg_freq_30_40 | |

**Figure 4.5:** Performance achieved by incrementally adding features selected by forward search.

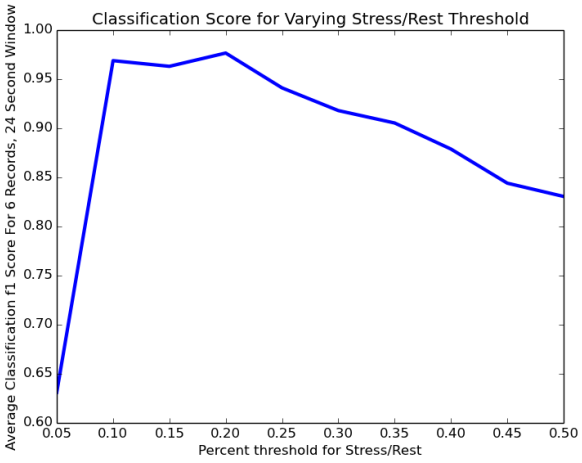



**Figure 4.3:** Performance with our best classifier (linear SVM, 24-second windows) at different cutoffs for the training window lengths.

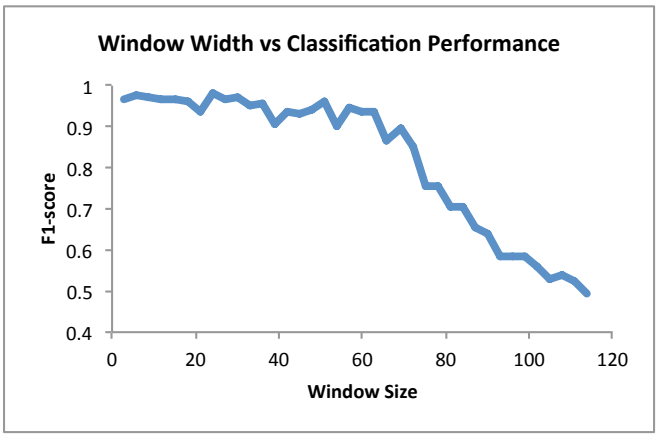



**Figure 4.4:** Classification accuracy ($F_1$ score) with varying window sizes.

Forward search (Figure 4.5) indicated that the majority of the classification performance was due to just a few top features: the ECG features in various bands, and the PNN50 feature (percentage of adjacent normal RR intervals that differed by more than 50ms). This indicates that immediate variation in heart rate is a good indicator of stress.

## 5 Conclusion

Using a combination of HRV and ECG features, we are able to predict whether the GSR is in the highest or lowest 20 percentile with 98% accuracy. Samples with less extreme GSRs can be predicted with 85% accuracy. This suggests that using consumer ECG devices, we would be able to predict whether or not the user is stressed with high confidence without requiring GSR measurement.

We draw the following conclusions.

- **HRV is not enough to reliably determine stress state.** Low-end consumer devices only transmit heartbeat timing and not the full ECG signal. The added information from ECG spectral features increased classification score from $F_1 = 0.78$ (with only HRV) to $F_1 = 0.98$ (with HRV and ECG spectral features combined).
- **ECG responds quickly to stress stimuli.** We were expecting that the GSR response might occur significantly before or after the ECG response, but offsetting the windows only decreased classification score.
- **ECG captures subtle heart activity variations.** There is valuable additional data in the raw ECG waveform, perhaps due to intrabeat variations in wave morphology or due to increased movement-induced noise when subjects were stressed. If the wave morphology itself changes, follow-up studies

should consider incorporating time-domain features of the QRS complex. Each ECG cycle typically has five deflections (P, Q, R, S, T), and meaningful physiological information might be derived from variations in the relative lengths and amplitudes of these sub-waves.

- **Stress is continuous, but a binary classification can often be sufficient.** Stress is an inherently continuous (and likely multi-dimensional) variable, so this binary classification as "stress" or "rest" is a simplification that would be appropriate in applications that would have to take a specific action in response to a certain stress threshold level. This level could be varied by changing the weighting of classes in the SVM.

### 5.1 Future Work

Other approaches that might be considered include dimensionality reduction using PCA.

A deeper qualitative understanding of which features are most important could help us discover more features and better understand the biological processes related to activation of the sympathetic nervous system.

We might consider extracting different HRV features: Barbieri et al. propose modeling the heartbeat intervals as an inverse Gaussian process, and describing the HRV using the parameters of this process [8].

We might also try including finer-grained bins of the ECG spectrum. We recently learned that most of the information content in ECGs is contained in the 0-40 Hz band, so more resolution in that area might improve results. Other researchers have suggested that other time-frequency analysis algorithms (such as the Wavelet transform) may improve discrimination of ECG data [9].

Another dimension to explore would be to distinguish between positive and negative emotional stress (eustress and distress, respectively)—we currently make no distinction. Our data set probably consisted of mild distress, a neutral state (rest), and few incidences of eustress, because it was all collected during automobile driving. Some researchers have shown that in certain cases heart rate increases more for positive arousal than for negative arousal [10].

### 5.2 Applications

Measuring stress with ECG has diverse applications. Psychology experiments measuring stress levels would no longer require a subject's hands to be free, would require less intrusive equipment, and could reduce cost. One consumer application is a "stress alert" system that could be worn at all times of day. Most current biofeedback systems rely on hand GSR, which is cumbersome and cannot be worn continuously.

For a consumer application, we would consider using an online and incrementally adaptive learning system. Initially, with only generic training data, we may achieve classification accuracy of approximately 85% (the accuracy we achieve when pooling different subjects' records). User feedback could be used to generate further training examples for an incremental online learning algorithm. We plan to test our models for stress prediction and binary stress classification with real-time data from a Bluetooth ECG device.

## References


[1] Lawrence R Murphy. Stress management in work settings: a critical review of the health effects. *American Journal of Health Promotion*, 11(2):112–135, 1996.

[2] Franklin Stein. Occupational stress, relaxation therapies, exercise and biofeedback. *Work: A Journal of Prevention, Assessment and Rehabilitation*, 17(3):235–245, 2001.

[3] S. R. Vrana. Emotional modulation of skin conductance and eyeblink responses to startle probe. *Psychophysiology*, 32(4):351–357, Jul 1995.

[4] Jennifer A Healey and Rosalind W Picard. Detecting stress during real-world driving tasks using physiological sensors. *Intelligent Transportation Systems, IEEE Transactions on*, 6(2):156–166, 2005.

[5] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[6] Sansanee Boonnithi and Sukanya Phongsuphap. Comparison of heart rate variability measures for mental stress detection. In *Computing in Cardiology, 2011*, pages 85–88. IEEE, 2011.

[7] Tracy Chou, Yuriko Tamura, and Ian Wong. Detection of atrial fibrillation in ecgs. *CS 229 Final Class Projects*, 2008.

[8] R Barbieri, EC Matten, and EN Brown. Instantaneous monitoring of heart rate variability. In *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, volume 1, pages 204–207. IEEE, 2003.

[9] AKM Fazlul Haque, Md Hanif Ali, M Adnan Kiber, and Md Tanvir Hasan. Detection of small variations of ecg features using wavelet. *ARPN Journal of Engineering and applied Sciences*, 4(6), 2009.

[10] Peter J Lang, Mark K Greenwald, Margaret M Bradley, and Alfons O Hamm. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30(3):261–273, 1993.

[11] W Picard and Jennifer A Healey. Wearable and automotive systems for affect recognition from physiology. 2000.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.