Arousal Detection for Biometric Data in Built Environments using Machine Learning

Abstract

This paper describes an approach using wearables to demonstrate the viability of measuring physiometric arousal indicators such as heart rate in assessing how urban built environments can induce physiometric arousal indicators in a subject. In addition, a machine learning methodology is developed to classify sensor inputs based on annotated arousal output as a target. The results are then used as a foundation for designing and implementing an affective intelligent systems framework for arousal state detection via supervised learning and classification.

1 Introduction

1.1 Goals

In this paper, we propose using machine learning classification techniques such as support vector machines (SVMs), general linear mixed models (GLMMs), logistic regression (LR), and artificial neural networks (ANNs) to demonstrate the viability of using automation and machine learning techniques in classifying biometric arousal state, as determined by domain expert annotation. We address the task of learning a classification-based signal identification model for arousal response from multichannel sensor data produced in a built environment. This task entails a need for ground truth annotations, for which we develop a rating scale based on definitions given by neurobiological domain experts to support annotations by such experts.

The motivating goal is to develop an intelligent system to both classify and predict biometric arousal state, automating a process that is traditionally performed by human experts in both physiometric signal identification and environmental sciences. Unique aspects of this approach include using machine learning on location-aware time series data (the topic of this paper) and potential future work on multisensor integration using a range of wearable sensors together with images of the built environment to incorporate visual stimuli.

1.2 Limitations of Existing Work

Traditionally, it has been necessary to measure physiometric arousal indicators such as temperature, galvanic skin response (GSR), and heart rate in humans by relying on high precision laboratory equipment. The subject in an experiment will often be required to have sophisticated equipment attached to them in order to monitor and collect data from them. In addition, this information is often collected from the subject in a controlled environment such as walking or running on a treadmill while data is being recorded. The advent of wearables such as the Empatica E4, Polar, and Garmin Vivosmart 3 provides researchers with the ability to conduct new experiments such as measuring physiometric arousal indicators induced by a subjects urban environment. These sensors include high quality pedometers, optical heart rate omnitor, accelerometers, barometer, GPS, and GSR allowing the researcher to collect high quality data in a non-intrusive manner. Geolocation sensors and chronometers on some wearables now enable the collection of geospatial data for spatiotemporal analytics.

Despite the advances in measurement technology, the current state of the field for built environments still relies completely on human defined expert annotation. In other words, the current state of the field is notable in its absence of using automation and machine learning approaches for classification and prediction. We seek to explore how to fit a classifier model that generalizes over individual routes to impute arousal state by classification in a manner consistent with the state defined, and identified in historical data, by a human expert annotator. In addition, the full potential of real-time data collection, annotation, and prediction of arousal given a subjects environment by using the inputs collected by wearable devices as described above has not been sufficiently explored. As such, fitting machine learning algorithms and models, to address the problems posed by this new research domain, has not been sufficiently explored.

1.3 Objectives and Significance

Our desire is to demonstrate that our experimental approach will show promise in the the classification and detection of arousal. By doing so, we hope to be able to begin constructing a machine learning framework for [developing] a predictive intelligent system in future work. In addition, we desire to determine which machine learning classification approaches are appropriate over others.

This paper proposes significance by making two novel contributions to state of the field. The first is to demonstrate the viability of machine learning algorithms being an appropriate venue to fit data in a built environment scenario. The second is to motivate future work by showing several open questions suggested by the findings in this paper.

Here, *built environment* is defined as the human developed space that is comprised of where people live, work, and recreate on a daily basis [Roof and Oleru, 2008]. The second is since the application domain field relies on expert annotation of arousal, developing an effective model for classification and prediction will demonstrate the viability our of approach and provide a baseline for further research into intelligent systems.

We contend that arousal detection using biometric, environmental variables, and neurophyscologist annotation is an area that is relatively unexplored as a machine learning task in built environments. Exploring models and training them to fit data to detect arousal using the above features is possible.

Concretely, this paper proposes the goal of detecting expert-annotated arousal or arousal measurement by classifying a stream of observations as belonging to an arousal event or not.

1.4 Central Hypothesis

We have two hypothesis we wish to test. The first is discrimination of the data. That is, can we fit a model using machine learning algorithms to the data. The second is to consider the accuracy and AUC (area under the curve) of SVM, ANN, Logistic, and GLMM on the validation sets on 3 and 4 fold cross validation. We train our models and then compare the mean accuracy and AUC of our tests. The conservative Mann-Whitney test which makes no parametric distributional assumptions is used. We formalize our hypothesis as follows:

$$H_0 : \mu_{A1} - \mu_{A2} = 0$$
$$H_A : \mu_{A1} - \mu_{A2} \neq 0$$

Here $A1$ and $A2$ denote distinct machine learning algorithms such as SVM, ANN, Logistic, or GLMM. The test uses $\alpha = 0.05$ and results are shown below in the evaluation strategy and results section.

1.5 Approach

Evidence suggests linkages between the physical environment and its influence on mental health, well-being and human health [Evans, 1984; Kuo *et al.*, 1998; Evans, 2003; Abraham *et al.*, 2010; Berman *et al.*, 2012]. However, there is much to learn about how particular design characteristics, natural elements, architecture and planning play a role in influencing well-being. Fortunately, there are indications that natural elements do improve mental [Ulrich, 1981; Parsons *et al.*, 1988; Parson *et al.*, 1998] and physical health (new research LURP). With global urbanization, the pressures for development and density are increasing. At times, urban development places pressure on the availability of outdoor amenities and recreational spaces because these may be seen as less valuable than the proposed built-infrastructure, but are critically important [Groenewegen *et al.*, 2006; Maas *et al.*, 2006]. Unfortunately, the replacement of nature with built infrastructure may negatively impact public health at-large [Jackson, 2003], leading to a greater risk of suffering from conditions such as stress and mental fatigue [Ulrich, 1981; 1983; Kaplan and Kaplan, 1989; Kaplan, 1995; Parson *et al.*, 1998] .

Growing our knowledge of how design and urban form influence mental health is a critical issue in the 21st century. With new technologies and methods, we are now able to investigate relationships between the built-environment and human affective response in order to ascertain how design and planning of urban spaces may influence well-being. Research conducted by Ulrich [1991] and Tsunetsugu et al. [2013] suggests that there are strong physiological responses (e.g. reductions in heart rate) to viewing nature. Whereas their research was conducted in laboratory settings with discrete or short-term data, new machine learning techniques and wearable sensors [Poh *et al.*, 2010] offer a unique opportunity to investigate affective responses to elements of urban form and assess the extent to which these elements influence long-term mental health and stress.

The experiment relies on the Empatica E4 wearable and consists of 12 subjects who volunteered to participate. The experiment was conducted by placing an Empatica E4 sensor on each individual subject. This sensor measures temperature, galvanic skin response, heart rate, time, and geospatial position.

Each individual walked a predetermined route in a Manhattan, Kansas urban environment divided into a series of zones selected to reflect a specific urban setting. Examples include a dark alley, poorly lit street, well lit sidewalk, and calming park areas. For the control, the user was asked to sit and calmly walk from a predetermined starting point at a hotel to the beginning of the route. This two-minute period toward the experimental route provides the baseline data for heart rate, temperature, and galvanic skin response. Each participant after the experiment was given a survey and rated the perceived safety of each zone. The data outside of zones in the survey are not rated by participants and therefore receive a zero arousal score by default.

The data has been cleaned and processed and is organized by participant ID. The data was processed and now has fields such as zone, ratings of zones, 30 second window giving heart rate and standard deviation. The classification target is a binary variable annotated by Dr. Greg Norman.

The data has been trained on several applied machine learning techniques such as SVM, ANN, LR, and GLMM. Performance is measured by using accuracy and AUC.

2 Background and Related Work

2.1 Related Work

Research on estimating arousal using wearable technology has its roots in the late 1990s with the dawn of wearable computing. In the last few years there has been growing interest in this area due to the increasingly abilities and capabilities of wearable and mobile computing. This area of research is still nascent and specific signal identification, pattern detection, and prediction methodologies not only constitute a new approach but have yet to be applied and refined for many use

cases, such as arousal estimation and prediction in built environments.

Recent work has begun to consider the role that machine learning can play in measuring arousal using mobile and wearable technology. Specifically, using supervised learning methods such as linear regression and support vector machines (SVM) to classify arousal [Hernandez Rivera, 2015]. In addition, work has been done in developing a user tailored advice system feedback loop using wearable and mobile computing for arousal intervention regarding sleep, diet and exercise habits. The work was statistical in nature and presents an opportunity to build upon it using more sophisticated machine learning approaches. Users with the higher levels of arousal measured reported appreciating the intervention feedback [Sano *et al.*, 2015]. Sano and Picard [2013] have used binary classification with correlation analysis to determine physiological or behavioral markers for arousal using wearables and mobile computing. The study showed that higher levels of arousal were correlated with activity level and screen on/off patterns [Sano and Picard, 2013]. Feasibility studies have shown that it is possible to classify and predict panic attacks using wearables and mobile computing using intelligent systems [Rubin *et al.*, 2015]. The United States Army is showing interest in using wearables to collect real time information from the soldier [Hoyt and Karl, 2016]. Future studies will likely consider arousal detection and prediction as well. Very recently, industry has been developing products which claim to detect stress in users as well by using variable heart rate [Lisanti, 2017].

Despite the recent work in measuring arousal using wearable technology, there are a plethora of methodologies, measurements, and instruments for measuring arousal that lead to inconsistent results [Lutchyn *et al.*, 2015]. Clearly, there is a lot of work that remains to be done in general and in built environments.

2.2 Heart Rate and Arousal Estimation

Heart rate (HR) and electrodermal activity (EDA) data were used to generate predictions about the affective state of each participant as they walked through predetermined zones that ostensibly varied on dimensions of nature and urban along with mild threat and safety [Thayer and Lane, 2009]. Data were aggregated over 30 second segments for each participant. Additionally, all individuals completed a baseline session prior to the walking component of the experiment in order to determine within-participant change in physiological activity during the walk [McEwen, 2007]. In an attempt to predict the general affective state of the participant while they walked through different zones, we normalized the HR and EDA data within each participant and evaluated the change in these signals within each zone. Zones that were associated with the largest deviations from baseline were labeled stress and zones that were associated with minimal change were labeled as no-stress. For the purposes of this study, the stress vs. no-stress distinction was determined using a threshold of 2 + standard deviation change from the baseline condition for each participant.

2.3 Support Vector Machines

Support vector machines (SVMs) are a standard approach in solving classification problems. They are common supervised learning tools that fall under the of *large margin methods* (for minimizing the statistical risk of decision surfaces) and *kernel methods* for rendering implicit those mappings designed to change the learning representation by reformulating the instance space.

The problem of estimating model parameters is specified as a convex optimization problem. That is, any local solution will also be the global optimum [Bishop, 2006]. The SVM is a model which maps all observations upon a plane and divides them with a linear separator and margin. The linear separator and margin are what separates the observations into two classes. In other words, the decision boundary is chosen to be the one for which the margin is maximized [Murphy, 2012].

2.4 Multilayer Perceptrons

Multilayer perceptrons (MLPs) are a type of feedforward artificial neural network (ANN) which, like SVMs, are extremely popular as inductive learning representations. They are currently an area of intense study as an essential component of deep learning. A function learns from inputs and adjusts weights using a hidden layer, which in turns uses an activation function to simulate the threshold and action potential of a simulated neuron. As biologically-inspired models, MLPs and other feedforward ANNs provide flexibility as to the type of nonlinear activation functions, pooling functions, interlayer connectivity, overfitting control methods, and other representational properties that enable them to function as autoencoders in deep learning.

2.5 Logistic and General Linear Mixed Models

The most common classification method for linear method is logistic regression. The classification is given as a posterior probability which relies on a logistic sigmoid acting on a linear function [Bishop, 2006; Cox, 1958].

General linear mixed models (GLMMs) are an extension of standard general linear model to include fixed effects, random effects, and autocorrelation. The unique aspect of GLMM is that the response variable can come from different distributions besides the normal distribution. In addition, rather than directly modeling the responses directly it is common to apply the data to link function. Concretely, a general linear mixed model may be described as:

$$y = X\beta + Z\gamma + \epsilon$$

where y is a $N \times 1$ column vector, X is a $N \times p$ matrix of p regressor variables, β is a $p \times 1$ column vector of fixed-effect coefficients, Z is $N \times q$ matrix of q random variables, γ is vector of random effects, and ϵ is a $N \times 1$ column of errors not explained by the model.

It is not tenable to compute the exact likelihood function for GLMMs. Breslow and Clayton provide an algorithm to approximate the likelihood function [Brewslow and Clayton, 1993]. This approach is known as partial quasi likelihood (PQL) and approximates high dimensional integration using a Laplace approximation.

Consequently, the GLMM model has been developed to address data that is binary and also has autoregressive features.

3 Methodology

3.1 Data Preparation

The data has been cleaned and prepared. It is organized by participant id. The study has a window size of 30 seconds, with a mean heart rate and standard deviation for that period. The data has zone rating which is the post survey results taken by each participant to answer questions about how they feel about zones of interest they walked through. In addition, an expert has annotated the data for arousal using a binary variable 1 for arousal and 0 for no arousal. The data is show below in a table.

Table 1: Schema of Data

Variables
Time
Participant
Temperature
Average EDA
Speed
Latitude
Longitude
Heart Rate (HR)
Ratings
Number Street Lights
Max Road Width
Tree Frequency
Question Zone
Walk Score
Total Vegetation Sqft
Arousal

The above table consists of the schema for the experimental data. As shown above, each row has the time, participant id, average eda, walking speed, latitude, longitude, heart rate, and the likert rating of the question zone denoted by the variable ratings. In addition, we have variables we collected from the built environment such as number of street lights, max road width, what zone they were in during the walk (called question zone), a score of the walk annotated by a professional landscape expert, total vegetation of the area in sqft, and finally the variable we wish to estimate which is the presence of arousal or not.

It is important to note that data for participants 2,3,4,12, and 16 were removed since the participants either did not follow the directions appropriately or there was no accompanying zone rating data recorded. The data was annotated by Dr. Greg Norman, an expert in neuropsychology.

4 Experiment Design: Evaluation Strategy

This section discusses variables present in the building custom and novel machine learning models in this paper. The evaluation strategy is to use the core variables in the data that could help explain arousal in the data without introducing bias or correlation. Correlation is an issue with this data set. For example, lat/long, zone and zone ratings are correlated. Thus, only zone ratings is considered. In addition, the quality of gsr was not certain and ignored. Therefore, the core input variables are shown below:

4.1 Model Selection and Discrimination Strategy

Our approach for the detection and classification of arousal can be understood by performing model selection and discrimination on the data.

We began simply with a full model and observed which variables were statistical significant. We then removed, one at a time, variables not statistically significant or necessary to fit the data. The small sample size of participants and data for each participant due to smoothing the data per 30 second window should be noted. First, it is believed interpretation of the models parameters is not as important as demonstration that a model can be fitted. Second, the variables in the model fitted are of interest to us in motivating subsequent studies. We discuss the results of this strategy in section 5.2.

4.2 Cross validation Strategy and Calibration Strategy

In order to access the accuracy of detecting for arousal, the data is divided into training and testing data sets. Specifically, leave one out, 3 fold, and 4 fold cross validation has been conducted. We chose four and three fold cross validation only taking into account the small sample size and wanting to emphasize the possibilities, but introductory nature of the analysis to inform future studies. Since our data is longitudinal and organized by participant ids, we name and dived the data into sections based on participant id. For example, for the first fold of 4 fold cross validation we trained the model on participants 1, 5, 6, 7, 8, 9, 10, and 11. We then accessed accuracy on participants 13, 14, 15, and 17.

Figure 1: 4 Fold Cross Validation

Fold	Train	Test
1	1,5,6,7,8,9,10,11	13,14,15,17
2	1,5,6,7,13,14,15,17	8,9,10,11
3	8,9,10,11,13,14,15,17	1,5,6,7

The data is divided into training and testing data sets by participant id using 3-fold cross validation as follows:

Figure 2: 3 Fold Cross Validation

Fold	Train	Test
1	1,5,6,7,8,9,10,11,13,14,15	17
2	1,5,6,7,8,9,10,11,13,14,17	15
3	1,5,6,7,8,9,10,11,13,15,17	14
4	1,5,6,7,8,9,10,11,14,15,17	13
5	1,5,6,7,8,9,10,13,14,15,17	11
6	1,5,6,7,8,9,10,13,14,15,17	11
7	1,5,6,7,8,9,11,13,14,15,17	10
8	1,5,6,7,8,10,11,13,14,15,17	9
9	1,5,6,7,9,10,11,13,14,15,17	8
10	1,5,6,7,8,10,11,13,14,15,17	7
11	1,5,7,8,10,11,13,14,15,17	6
12	1,5,6,7,8,10,11,13,14,15,17	5
13	1,6,7,8,10,11,13,14,15,17	1

The models were then trained and tested on SVM, ANN, GLM, and Logistic Regression. The small sample size of the participants and data collected due to smoothing by a 30 second window per each participant is noted and emphasized here. This analysis is preliminary, but believed it is still worth merit to investigate model accuracy and AUC to inform subsequent studies.

5 Experiment Design: Results

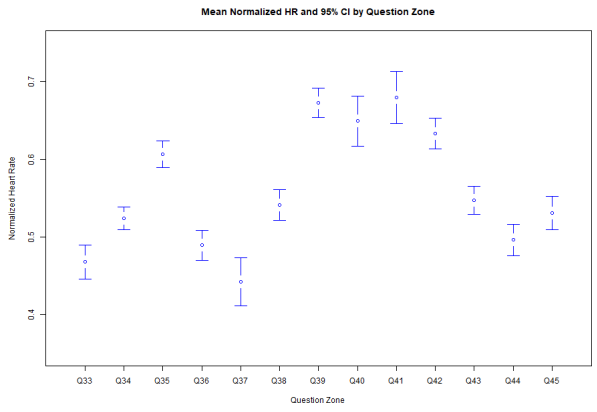
5.1 Normalized Heart Rate by Question Zone

We wanted to examine the raw heart rate data of participants as varied by the question zones to see if there were any notable differences. The data was normalized as follows:

$$normhr(x_i) = \frac{x_i - minhr(x)}{maxhr(x) - minhr(x)}$$

where x denotes the participant and i an arbitrary observation in the data. Please note that $minhr(x)$ and $maxhr(x)$ are global extrema based upon all heart rates for the participant. We used the normalized heart rate and calculated the mean for all participants for each zone given the 95% confidence interval.

Figure 1: Mean Normalized HR and 95% Confidence Interval by Question Zone



Looking at the figure above, one can note that there are differences between the normalized HR and that these mean heart rates tend to cluster or group. That is, the normalized mean HR in $Q33$ to $Q38$, $Q39$ to $Q42$, and $Q43$ to $Q45$ tend to group around a particular mean normalized HR while differing with normalized HR in other groups. We conclude that there are differences in normalized mean HR between question zones for participants.

5.2 Model Selection and Discrimination

The scope of our hypothesis, given the small sample size, was to assess the ability of machine learning models to fit the data in this context. We fitted full models for SVM, ANN, Logistic, and GLMM models. Our approach is standard, we removed one variable at a time in model specification until we came down to heart rate, temperature, and speed. The misclassification errors for all algorithms did not change that greatly, so we opted to keep speed in the model given the preliminary nature of the analysis. We assessed each algorithm on heart rate, temperature, and speed. The p-values for logistic and glmm can be seen below:

Figure 3: Estimated Model Variable P-Values		
Term	LR	GLMM
HR	0.000	0.000
Temp	0.000	0.014
Speed	0.500	0.4128

This is an interesting result, but not entirely surprising. Our expert annotation was primarily based on smoothing of the heart rate. Thus, it is not surprising that HR would be a part of the final model. It is interesting that temperature, another biometric measure, was also statistically significant. The misclassification errors for the machine learning algorithms is given below:

Figure 4: Misclassification Error Rate				
Term	SVM	ANN	LR	GLMM
Error Rate	0.092	0.150	0.137	0.142

It is interesting to note that in model selection, usually the problem is from over fitting the models, but that is not the case here. Since the data is binary in nature, we wish to conduct future studies that both smooth over smaller windows (5 second, 10, and 15 second intervals instead of 30 seconds in this study) and collect continuous annotation data of stress. That is, instead of taking post survey questionnaires and relying only on expert annotation of stress we provide users with a sensor that they can input their stress levels continuously in real time.

5.3 Model Calibration

We compare SVM, ANN, LR, and GLMM. Specifically, we compare accuracy and AUC. The conservative Mann-Whitney test revealed no statistical differences between the algorithms compared in accuracy and AUC. Consequently, we did not report mean and confidence intervals in the figures below. Although the trends observed were not statistically significant, the models performed considerably well given the small sample sizes. The results below are for 4 fold CV. Please note columns 1, 2, and 3 refer to folds. Please see below:

Figure 5: 4 Fold CV Accuracy			
Algo	Fold 1	Fold 2	Fold3
SVM	0.756	0.756	0.832
ANN	0.876	0.821	0.744
LR	0.892	0.821	0.824
GLMM	0.892	0.902	0.816

The accuracy scores are extremely close to each other and no statistically significant differences were detected between them. We also report AUC scores as follows:

Figure 6: 4 Fold CV AUC			
Algo	Fold 1	Fold 2	Fold 3
SVM	0.756	0.756	0.832
ANN	0.876	0.821	0.744
LR	0.892	0.821	0.824
GLMM	0.892	0.902	0.816

The AUC scores were also very similar to each other across the algorithms under consideration. The differences between the algorithms were not statistically significant. The results below are for 4-fold CV. We compare SVM, ANN, and GLMM accuracy and AUC results. Please note columns 1, 2, 3, and 4 refer to folds.

Figure 7: 4 Fold CV Accuracy				
Algo	Fold 1	Fold 2	Fold 3	Fold 4
SVM	0.693	0.833	0.956	0.739
ANN	0.836	0.802	0.945	0.652
LR	0.867	0.802	0.956	0.760
GLMM	0.887	0.895	0.956	0.739

Like 3 fold cross validation, the results are similar across algorithms. In addition, there are no statistical results observed. The AUC scores for 4 Fold CV are as follows:

Figure 8: 4 Fold AUC				
Algo	Fold 1	Fold 2	Fold 3	Fold 4
SVM	0.407	0.602	0.804	0.681
ANN	0.871	0.260	0.405	0.537
LR	0.864	0.864	0.981	0.591
GLMM	0.866	0.846	0.990	0.595

It is of interest to note there are some decreases in AUC in SVM and ANN. These differences were not enough to be statistically significant. Nevertheless, it is of interest for subsequent studies to explore the potential differences between linear models and other machine learning classification algorithms. The algorithms did considerably well given the small sample sizes, but we believe further study is merited with more environmental variables should be included in subsequent studies.

6 Summary

This initial experiment has yielded preliminary results that demonstrate how it is feasible to learn classification-based models of arousal state from a combination of biometric data and built environment data. We have shown the potential for these algorithms to have tolerable misclassification errors and demonstrated the potential for prediction by promising AUC scores. We note the sample size is small and that statistical inference cannot be made. In addition, we also convey that the results did not allow us to differentiate which methods might be better. Nevertheless, we believe the potential for machine learning to be applied in this problem domain has been sufficiently suggested and merits further research.

6.1 Limitations and Open Questions

This research has generated several open questions during our investigation. The sample size of the data was limited to 13 participants. It is desirable that future work exceed this number and seek as many participants as possible to reduce variability and increase the potential for statistical inference.

Results have shown that machine learning algorithms can be trained to fit biometric built data and that while inconclusive, AUC curves are promising. Unfortunately, the small sample size means that statistical comparisons between the machine learning algorithms is inconclusive both in fitting and validating the data. It is an open question which methods will give the researcher better fit and prediction. Concretely, whether linear classification models being compared to neural classification algorithms.

The expert annotated data was binary. Data was averaged by 30 seconds to provide a mean heart rate and standard deviation to help determine if a arousal event was observed. It is unknown how a smaller window might have influenced the machine learning algorithms.

Results shown that the models selected favored heart rate, which correlated unsurprisingly with the prediction target of arousal. It an open question what is the best prediction target for the estimation of arousal given a built environment. Specifically, it is of interest to determine whether a discrete target variable or continuous variable target will lead to better machine learning results and statistical inference.

It is unknown what are the essential variables that could potentially aid building a machine learning model both fits the data and predicts future arousal in built environments well. Therefore, it is of interest to collect as many variables in a participants environment as possible. We believe there is merit in also exploring how to best collect this environmental data of their built environment and represent it to the the machine learning algorithms.

The ability to detect an arousal event in a user given their built environment presents interesting and new possibilities in affective computing research. There are many questions about which is the most appropriate approach for sensor fusion, model building, and how to build applications which could affectively respond to a user based on their arousal state.

6.2 Future Work

In our continuing work on developing this multisensor approach to arousal state detection and prediction, we plan to perform another experiment with as many participants as possible. Our experiment design involves users being supplied with mobile sensors that they can use to annotate their perceived arousal state given their built environment in real time. We intend to explore the performance of machine learning algorithms in a continuous prediction variable target scenario. At the same time, we wish to observe as many variables as we can about the environment they are in, and collect as much

biometric data from the user as possible. In order to develop a model that can fit the data well and be used for prediction, we will systematically explore machine learning representations (especially loss functions and hyperparameters) and algorithms to determine those most appropriate to this problem. These specifically include linear methods suitable for discrete variables or continuous spatiotemporal domains. Our overall goal is to devise a machine learning based methodology for both detection and prediction of arousal events in a built environment that can provide affective response feedback to a user.

Acknowledgments

We acknowledge the gracious and kind assistance, comments, and feedback provided by Dr. Weixing Song during the course of this experiment.

References

[Abraham *et al.*, 2010] A. Abraham, K. Sommerhalder, and T. Abel. Landscape and well-being: a scoping study on the health-promoting impact of outdoor environments. *International journal of public health*, 55:59–69, 2010.

[Berman *et al.*, 2012] M. G. Berman, E. Kross, K.M Krpan, M. K. Askren, A. Burson, P. J. Deldin, S. Kaplan, L. Sherdell, I. H. Gotlib, and J. Jonides. Interacting with nature improves cognition and affect for individuals with depression. *Journal of affective disorders*, 140:300–305, 2012.

[Bishop, 2006] C Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, New York, 2006.

[Brewslow and Clayton, 1993] N. E. Brewslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1993.

[Cox, 1958] D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society*, pages 215–242, 1958.

[Evans, 1984] G. W. Evans. Environmental stress. *CUP Archive*, 1984.

[Evans, 2003] G. W. Evans. The built environment and mental health. *Journal of Urban Health*, 30:536–555, 2003.

[Groenewegen *et al.*, 2006] P. P Groenewegen, A. E. Van den Berg, S. De Vries, and R. A. Verheij. Vitamin g: effects of green space on health, well-being, and social safety. *BMC public health*, 6:149, 2006.

[Hernandez Rivera, 2015] J Hernandez Rivera. *Towards wearable stress measurement (Doctoral Dissertation)*. MIT Press, Cambridge, Massachusetts, 2015.

[Hoyt and Karl, 2016] R. Hoyt and P. Karl. The future of wearable tech. *US Army. Retrieved from https://www.army.mil/article/161761/*, February 2016.

[Jackson, 2003] R. J. Jackson. The impact of the built environment on health: an emerging field. *American Public Health Association*, 2003.

[Kaplan and Kaplan, 1989] R. Kaplan and S. Kaplan. The experience of nature: A psychological perspective. *CUP Archive*, 1989.

[Kaplan, 1995] S. Kaplan. The restorative benefits of nature: Toward an integrative framework. *Journal of Environmental Psychology*, 15:169–182, 1995.

[Kuo *et al.*, 1998] F.E. Kuo, M.Bacaicoa, and W.C. Sullivan. Transforming inner-city landscapes trees, sense of safety, and preference. *Environment and Behavior*, 30:28–59, 1998.

[Lisanti, 2017] J. Lisanti. Garmin vivosmart 3 review: Testing out the tracker's stress monitoring, rep counter, and more. *Sports Illustrated. Retrieved from https://www.si.com/edge/2017/05/15/garmin-vivosmart-3-review-fitness-tracker*, May 2017.

[Lutchyn *et al.*, 2015] Y. Lutchyn, P. Johns, M. Czerwinski, S. Iqbal, G. Mark, and A. Sano. Stress is in the eye of the beholder. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction*, pages 119–124. IEEE, 2015.

[Maas *et al.*, 2006] J. Maas, R. A. Verheij, P. P. Groenewegen, S. De Vries, and P. Spreeuwenberg. Green space, urbanity, and health: how strong is the relation? *Journal of epidemiology and community health*, 60:587–592, 2006.

[McEwen, 2007] B. S. McEwen. Physiology and neurobiology of stress and adaptation: Central role of the brain integration. *Physiology Review*, 87:873–904, 2007.

[Murphy, 2012] K. P Murphy. *Machine Learning*. MIT Press, Cambridge, Massachusetts, 2012.

[Parson *et al.*, 1998] R. Parson, L. G. Tassinary, R. S. Ulrich, M. R. Hebl, and M. Grossman-Alexander. The view from the road: Implications for stress recovery and immunization. *Journal of Environmental Psychology*, 18:113–140, 1998.

[Parsons *et al.*, 1988] R. Parsons, L. G. Tassinary, R. S. Ulrich, M. R. Hebl, and M. Grossman-Alexander. The view from the road: Implications for stress recovery and immunization. *Journal of Environmental Psychology*, 18:113–140, 1988.

[Poh *et al.*, 2010] M. Z. Poh, N. C. Swenson, and R. W. Picard. A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE transactions on Biomedical engineering*, 57:1243–1252, 2010.

[Roof and Oleru, 2008] Roof and Oleru. Public health: Seattle and king county's push for the built environment. *J Environ Health*, 71:24–27, 2008.

[Rubin *et al.*, 2015] J Rubin, H. Eldardiry, R. Abreu, S. Ahern, H. Du, A. Pattekar, and D. G. Bobrow. Towards a mobile and wearable system for predicting panic attacks. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 529–533. ACM, 2015.

[Sano and Picard, 2013] A. Sano and R. W. Picard. Stress recognition using wearable sensors and mobile phones. In

Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, pages 546–552. ACII, 2013.

[Sano *et al.*, 2015] A. Sano, P. Johns, and M. Czerwinski. Healthaware: An advice system for stress, sleep, diet and exercise. In *Proceedings of the 2015 International Conference in Affective Computing and Intelligent Interaction*, pages 546–552. IEEE, 2015.

[Thayer and Lane, 2009] J. F. Thayer and R. D. Lane. Claude bernard and the heartbrain connection: Further elaboration of a model of neurovisceral integration. *Neuroscience Biobehavioral Reviews*, 33(2):81–88, 2009.

[Tsunetsugu *et al.*, 2013] Y. Tsunetsugu, J. Lee, B. J. Park, L. Tyrvinen, T. Kagawa, and Y. Miyazaki. Physiological and psychological effects of viewing urban forest landscapes assessed by multiple measurements. *Landscape and Urban Planning*, 113:90–93, 2013.

[Ulrich *et al.*, 1991] R.S Ulrich, Robert F. Sims, Barbara D. Losito, Evelyn Fiorito, Mark A. Miles, and Micheal Zelson. The view from the road: Implications for stress recovery and immunization. *Journal of Environmental Psychology*, 11:201–230, 1991.

[Ulrich, 1981] R. S. Ulrich. Natural versus urban scenes. *Environment and Behavior*, 13:523, 1981.

[Ulrich, 1983] R. S. Ulrich. Aesthetic and affective response to natural environment. *Behavior and the natural environment*, pages 58–125, 1983.