JMLR: Workshop and Conference Proceedings 86:35–51, 2018 2nd International Workshop on Artificial Intelligence in Affective Computing

Binary Classification of Arousal in Built Environments using Machine Learning

Heath Yates HLYATES@KSU.EDU *Biosecurity Research Institute*
Brent Chamberlain BRENTCHAMERLAIN@KSU.EDU *Kansas State University*
William H. Hsu BHSU@KSU.EDU *Kansas State University*

Abstract

The goal of this paper is to develop a methodology and model to classify and characterize the arousal state of participants in a built environment. Demonstrating this showcases the potential of developing an intelligent system capable of both classifying and predicting biometric arousal state. This classification process is traditionally performed by human experts. Our approach can be leveraged to take advantage of the diversity of real-time sensor data to inform the development of smart(er) environments to improve human health.
Keywords: List of keywords

1. Introduction

1.1. Goals

This paper proposes using a binary classification machine learning technique such as logistic regression (LR) to detect user annotated affect in a built environment. The work in this paper is an extension of the work begun in Yates et al. (2017b), but with some notable differences. First, we consider user annotation of perceived safety in a built environment as ground truth for the detection of arousal in a built environment. Second, this paper considers an additional 6 participants. Third, no aggregate window was used to smooth the data and instead, the biometric data were normalized by participant and by zone. In addition, new environmental and participant characteristics were considered in the analysis.

1.2. Objectives and Significance

Research suggests strong correlations and relationships between a physical environment and the influence on the physical and mental well-being of humans Evans (1984); Kuo et al. (1998); Evans (2003); Abraham et al. (2010); Berman et al. (2012). There is also evidence that there is a complex relationship between city living, urban upbringing, and the effect of neural stress in humans Lederbogen et al. (2011). Feelings of safety have been linked as a specific affect regulation system which that has relationships to depression, anxiety, stress, and self-criticism Gilbert et al. (2008). The authors assert that an affective computing approach may offer a way to generate more responsive environments that improve well-being of those who use them. Wearable sensors and the activities of those using them can be better understood using machine learning techniques to create a viable framework in unify the design of built environments with computer science. In other words, affective

computing can provide a new metric to measure this subjective experience of participants in built environments.

1.3. Evaluation Strategy

This paper examined the predictive accuracy of the models by fitting machine learning algorithms to the data. Specifically, comparing logistic (LR) to RF, SVM, and MLP. In addition, Area under the Curve (AUC) was considered. The conservative Mann-Whitney test compares the accuracy and AUC scores of LR to other algorithms. The test uses $\alpha = 0.05$. Since LR performed statistically as reasonably well as the other algorithms, different models were evaluated on LR using model fit and performance such as AIC and McFadden's pseudo R^2 score. Please refer to section 5.5 for further discussion.

2. Background and Related Work

2.1. Wearables

The affective computing approach of detecting arousal using wearable technology is still young, but has been around since the 1990s with the dawn of wearable technology. This trend is likely to accelerate given the rise of high quality commercially available technology, such as Empatica, Fitbit, Garmin and a variety of commercially available smart watches. Currently, Empatica is considered medical grade quality and fitbit has been FDA approved Garbarino et al. (2014); Erdmier et al. (2016). The sophistication of these biomedical wearable sensors is not only expanding the horizons on what is possible in experimental design and data collection, but also in health care and Internet of Things (IoT) Mertz (2016); Metcalf et al. (2016).

In this paper, two sensors were used to collect biometric data. The first was the Polar V800 that collected GPS data and heart rate (HR). Research has been conducted on validating the HR to measure RR intervals at rest Giles et al. (2016). The second was the Empatica E4 which collected electrodermal activity (EDA), HR, and temperature. Research has shown the Empatica E4 has excellent performance metrics for EDA, HR, and temperature McCarthy et al. (2016). The sensor has also been used in multimodal data collection experiments for mental stress monitoring Kye et al. (2017).

Machine learning has played an inference role in measuring arousal using commercial and laboratory grade wearable technology. Notable experiments focused on linear and classification machine learning algorithms Sano and Picard (2013). Binary classification has been explored to correlated physiological and behavioural markers for arousal using wearables Garbarino et al. (2014), and it is possible to classify panic attacks using wearables and mobile computing Rubin et al. (2015). In the authors' earlier work, an analysis was conducted on participants outfitted with an Empatica E4 and a Polar wearable sensor, and they walked through a built environment. After the experiment, they filled out a survey indicating their responses to the environment. The biometric data and survey results were interpreted by an expert and annotated for the presence of arousal or not. The results demonstrated the viability of machine learning in a built environment context to detect annotation Yates et al. (2017b,a).

2.2. Random Forests

A classic and very effective machine learning method for binary classification and small data sets is an algorithm called random forests (RF). This method is primarily based on the principle of bagging with random feature selection that adds diversity to the decision tree model. After the collection of trees a forest is generated, the model then uses a vote to combine the tree's predictions Robert (2014); Christopher (2016). The algorithm is known for being very good at accuracy, handling large and small datasets, and being used for giving estimates of what variables are important in the data Lantz (2015).

2.3. Support Vector Machines

The support vector machine (SVM) is a classic machine learning algorithm that derives its name from the idea that a hyperdimensional plane is compressed down to a plane which divides the hyperdimension into two spaces. In other words, the algorithm partitions the data groups of similar data through a process referred to as the maximum margin hyperplane ensuring the greatest separation between the two classes. The algorithm has found notable success in the field of bioinformatics, text, and even in the detection of security breaches Wang (2005). The decision boundary is chosen to be the one for which the margin is maximized Murphy (2012). This can be concisely described as $min\frac{1}{2}\|\boldsymbol{w}\|^2$ such that $y_i(\boldsymbol{w} * \boldsymbol{x_i} - b) \geq 1$ for all $\boldsymbol{x_i}$ where all data points must satisfy the given constraint for the given margins Christopher (2016).

2.4. Multilayer Perceptrons

One of the most classic and well-known artificial neural networks is the three layer feed-forward neural network or multilayer perceptron (MLP), and is the *de facto* standard in ANN topology Lantz (2015). In a classic three-layer multilayer perceptron, there are three layers referred to as the input, hidden, and output layer. As the name indicates, the input layer is for the variables we wish to feed into the algorithm. The hidden layer processes the signals from the input before they reach the output layer Christopher (2016). Training of the network occurs through backpropagation Rumelhart et al. (1986). Artificial neural networks with at least one hidden layer also have been shown to be a universal function approximator, which means they can be made to approximate any continuous function Hornik et al. (1989). In the context of this paper, artificial neural networks will be used for binary classification.

2.5. Logistic Regression

Logistic regression (LR) is a very important, powerful, and spatio-temporal machine learning algorithm that has its origins in statistical learning Friedman et al. (2001). The name logistic implies that the relationship is a binary categorical outcome. It is very useful in binary classification and it's connection to statistics renders it an attractive model for inference. Regression here refers to specifying the relationship between the binary classification predictor to be estimated and the several input variables specified to describe the relationship Robert (2014). Logistic regression is also formally refereed to as a generalized linear

model for binary data where the outcome to be estimated is either a probability between 0 or 1. Mathematically, we can describe it as follows Christopher (2016):

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \text{ for all } x_i \text{ such that } i = 1, 2, ..., n$$

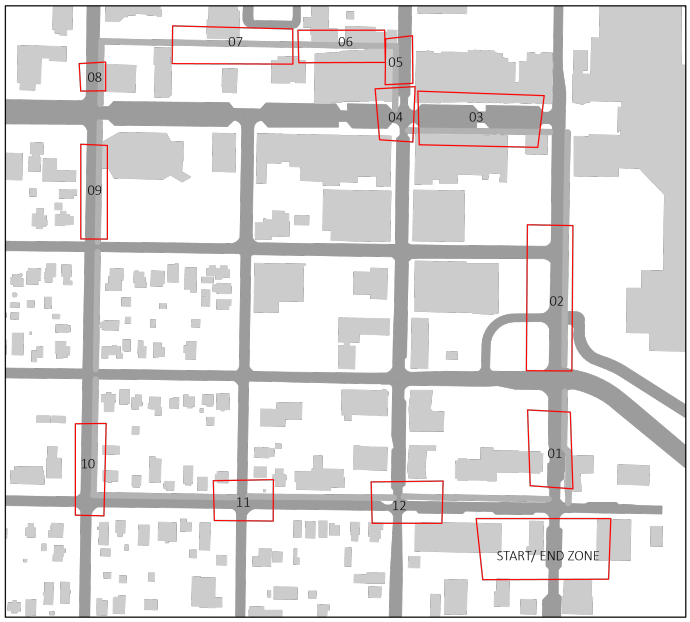
Let $\pi(x)$ denote the probability of the outcome such that $0 \leq \pi(x) \leq 1$ where the logit function $\log\left(\frac{\pi(x)}{1-\pi(x)}\right)$ has a binomial distribution. The β terms determine the rate of increase or decrease of logistic curve function. That is $\beta > 0$ implies an increase in $\pi(x)$ and $\beta < 0$ implies a decrease in $\pi(x)$. Otherwise the curve is flat Agresti (2002).

3. Experiment Design: Methodology

3.1. Approach

The experiment was designed and conducted in Manhattan, Kansas. Specific urban built environments were chosen by built environment domain experts based on environmental characteristics, such as walkability, number of trees, presence of grass, and the likelihood of the environment to invoke an arousal response. For the latter, examples include a darken alley, poorly or well lit streets, sidewalks, and calming park-like settings.

Figure 1: Urban Built Environment Annotated with Zones (Reprinted with permission from Whitaker (2018))



For this methodology, 12 distinct environmental zones were identified and geospatially delineated. The zones were named sequentially based on their location in the route as shown

above. In each zone, there are several urban built environment characteristics and features, such as trees, grass, pavement, buildings, and powerlines. While the authors acknowledge there are a wide number of possible environmental characteristics, only a few were evaluated for this stage of the methodology. The data set is comprised of 18 participants, which was collected from work done by earlier studies Parker (2016); Whitaker (2018). Each participant was fitted with a Polar V800 and an Empatica E4. Participants were then provided a map and asked to walk the designated route as indicated in the figure below.

Figure 2: Designated Route for Walk by Each Participant

The biometric data, such as HR and EDA, were normalized, and a baseline was established using the entire biometric data collected by participants. Please see the next section of this paper for further details. After the participant completed walking the route, the participant filled out a survey and rated the perceived safety of each zone. The survey was a Likert scale with 1 to 7, with 1 indicating an arousal event of feeling very unsafe versus 7 feeling very safe Parker (2016); Whitaker (2018). The tuples in the data outside of the zones in the survey were not rated and therefore not used beyond the normalization of HR and EDA. The data were cleaned, processed, and organized by participant ID and time. The user annotation of safety was filtered into a binary signal. Please refer to section 5.3.3 and 5.3.4 for further details. After the data were processed and cleaned, the data have been trained on several machine learning algorithms, such as LR, RF, SVM, and LR. This experiment used the standard methods provided by the R statistical language. For example, MLP has four units in the hidden layer, decay of 0.001, and max iterations of 1000. SVM has a cost of 100 and gamma of 1 Meyer (2004); Ripley et al. (2016). Please see section 5.4 and 5.5 for more details.

3.2. HR and EDA Normalization

Heart Rate (HR) and electrodermal activity (EDA) data were used as inputs in the machine learning algorithms to assist in generating estimates and predictions about the arousal state of each participant as they walked through each experimental zone. Research has shown that normalization of biometrics can be an useful and effective methodology in the detection of affective states Healey and Logan (2005); Healey and Picard (2005). The procedure was as follows. Normalization of HR and EDA was done by participant. The procedure for HR is as follows:

1. Find the top 10% and bottom 10% of the biometric data
2. Find the median of the top 10% and bottom 10% and denote it as top_{hr} and $bottom_{hr}$ respectively
3. Consider all HR tuples for a participant by zone, and find the lowest HR value that occurs in that particular zone. Denote this by $lowesthr_{zone}$
4. The normalized HR is calculated for each participant and zone as follows:

$$normhr = \frac{HR - lowesthr_{zone}}{top_{hr} - bottom_{hr}}$$

The procedure for EDA normalization was equivalent. This normalization approach provides an acceptable range and scale by participant while also removing the lowest outlier by participant per zone. This normalization and smoothing of the data provided the machine learning algorithms with normalized biometric data for each participant, removing extreme values and providing a scaled metric to better detect meaningful patterns in the data. Standard normalization for HR and EDA biometrics was performed as well, but the normalization above was deemed more useful both in model performance and better suited to the structure of the data based on methodologies and literature in affective computing.

3.3. Arousal Prediction Target

The detection of affect in the experiment was the estimation and prediction of ground truth affect or arousal by filtering the likert scale of perceived safety of in space from 1 to 7, where 1 is feeling very unsafe and 7 is feeling very safe, into a binary classification prediction target. For example, the likert scale implies a multinomial distribution and can make inference with machine learning algorithms not very tenable. Since the question at hand is to determine if core biometric and main environmental effects have an influence on arousal, it is quite reasonable to consider a binary classification target for easier inference and model building. In other words, the advantage of this approach is in parsimonious model building and inference by machine learning algorithms. The procedure for converting the likert scale to a binary classification can be described with the following simple procedure:

```
If(annotation < 5){
   annotation_{binary} = 1
}else{
    annotation_{binary} = 0
}
```

The likert scale and the boundary decision above forces the filter to unevenly spread the binary classification to denote 1 for arousal or unsafe affect versus 0 for no arousal or safe affect. Here, arousal describes a core affect more associated with feelings of being uncomfortable and unsafe.

3.4. Data Preparation

After normalization, the data were cleaned and organized and aggregated by participant. The Empatica E4 raw data was comprised of HR, EDA, and temperature. The raw Empatica data were the information that allowed the data once processed to be time stamped. The polar data contained timestamp and GPS information. They were processed and merged with environmental variables produced from Whitaker (2018). The environmental variables include number of street lights, number of trees, number of distinct grass spaces, walkability score, and the zone the participant was in. The characteristics were evaluated by aggregating the number of these features that intersected or were within 50 feet of the delineated zones. In addition, the polar data were tagged with a start and end times to indicate points where participants has walked for a block (to achieve a walking heart rate) and ended the experiment as shown in figure 5.2. These points helped to clean data that may have been messy beacuse of GPS inaccuracies as participants walked into and out of the building as a base for study logistics. The survey data were also cleaned and processed converting survey question numbers to the appropriate zone number, properly naming and numbering participants 1 to 18. In addition, biographical information was also verified and processed such as age, race, sex, body type, and urban background. The polar data were merged with the processed Empatica data by timestamp. The biometric HR and EDA data was normalized as described in section 3.3.2. The participant arousal response was filtered into a binary classification annotation target as described in section 3.3.3. Finally, the data were filtered to only include tuples for GPS coordinates that occurred in the experimental zones. The table below represents the schema of the cleaned and processed experimental data used for the analysis.

Table 1: Schema of Data

Variables
Participant ID
Normalized HR
Normalized EDA
Gender
Bodyshape
Urban Origin
Urban Preference
Study Area Familiarity
Exercise
Walkability
Number of Lights
Number of Trees
Number of Lines
Number of Points
Number of Grass
Number of Scrubs
Binary Annotation

Participant ID refers to the identification number assigned to the participant in the experiment. The normalized HR and normalized EDA is the processed HR and EDA biometric data as described in section 3.3.2. The demographic characteristics of the user is included in gender, bodyshape, and exercise. Urban origin and urban preference explore the origin where the participant grew up and if their preference is urban, suburban, or rural. Study area familiarity was a categorical variable that indicated the user's familiarity with experimental route. The environmental characteristics used in the data refer to the number of lights, trees, power-lines, points, grass, and shrubs in the experimental zones as determined by Whitaker (2018).

In addition to the scheme above, additional variables, such as skin temperature, age, and race, were explored but ultimately discarded from the data schema because they did not contribute to sufficiently explaining model building process according to the criteria outlined in the next section. Nevertheless, these variables still remain of interest in future work as the sample size of participants increases.

4. Experiment Design: Evaluation Strategy

This section outlines the criteria used in building custom machine learning models used in this experiment. At the heart of the criteria is the goal of building the most parsimonious model consisting of main effects or core variables that explain arousal in the data without introducing bias or correlation. Considering the issue of correlation as an example, zone is clearly correlated with environmental characteristics present in a zone, such as number of lights, trees, and grass. Therefore, this analysis considered models with only zone and those with the environmental characteristics. After an exploratory data analysis, the following models were chosen to be used in this experiment:

Table 2: Model Specification

Model	Variables
A	Normalized HR and EDA
B	Walkability, Number of Lights, Trees, Lines, Points, and Grass
C	All Variables Present
D	Normalized HR and EDA, Walkability, Number of Lights, Trees, Lines, Points, and Grass
E	Normalized HR and EDA, Zone, and Participant ID

In models A, C, D, and E normalized HR and normalized EDA were used. Model A can be thought of using only biometric signals to estimate user arousal. Conversely, model B relied only environmental variables. The full model is model C, which used all variables as specified in the data schema in section 3.3.4. Model D is the most nuanced model with biometric and select environmental characteristics chosen. The last model E relies primarily on zone and participant identification in conjunction with biometrics to explain arousal. The criteria used to compare these models is discussed in the next two sections. It is also important to note that this paper treated the environmental characteristics as factors for LR. The results are discussed in section 5.

4.1. Model Selection and Discrimination Strategy

First, an exploratory analysis was conducted on several models until the five models as specified earlier were chosen for a final comparison. The exploratory data analysis procedure was straightforward and proceeded by fitting the full model and then removing variables, one at a time, to see how the model would perform.

Second, criteria used for model selection and discrimination was placed upon accuracy and AUC scores. In addition, logistic regression was assessed with Akaike Information Criterio (AIC), Chi-Square Tests for fit, and McFadden's pseudo R-squared (pR^2). AIC is an estimator and score used for model selection between logistic models where $AIC = 2k - 2ln(\hat{L})$ such that L is the maximum value of the likelihood function for the model and k is the number of parameters in the model Neter et al. (1996). When comparing two models, the model with the minimum AIC score is to be chosen. Chi-square tests for fit are used to both measure if a null model or full model is appropriate. In addition, it can be useful in seeing if adding additional variables to the model is useful. Last, there is no strict measure of fit for logistic models like linear models have with R^2. Consequently, the Mcfadden's pseudo R^2 was devised for logistic regression as a measure of fit McFadden (1974). It can be succinctly described as follows:

$$pR^2 = 1 - \frac{ln(\hat{L}_{full})}{ln(\hat{L}_{intercept})}$$

In general, a score between 0.20 to 0.40 is considered the standard for a good fit and is the criteria used in this paper Hensher and Stopher (1979).

Third, this paper relies on properties of logistic regression and general linear models to make some interpretations of the model parameters in logistic regression. It is important to note that the interpretation is to be explanatory in nature and not necessarily predictive in nature at this time, but the results should be useful in future work.

4.2. Cross-validation Strategy and Calibration Strategy

This paper implements a cross validation strategy that focuses on the participants. That is, leave one out (LOOCV), leave 2 out (2FCV), and leave 3 out (3FCV). We briefly elaborate. For LOOCV, we train on all participants except labling one for testing and validation. This gives us 18 folds, one for each participant. For 2FCV, we train on 16 participants and validate on two. This gives us 9 folds. For 3FCV, we train on 15 and validate on 3 participants. This gives us 6 folds. The merit of implementing cross-validation is allowing a more comprehensive picture of model fit and potential for prediction to emerge.

The models as specified in the beginning of section 4 according to criteria outlined in section 4.4.1 were then trained and tested on LR, RF, SVM, and MLP. In addition, two pathological or naive models that predicted either all arousal or none were also fitted to the training and testing data above.

5. Experiment Design: Results

5.1. Comparisons

In this section, LR model D's accuracy is compared to the algorithms RF, SVM, and NN model D performance using the Mann-Whitney-Wilcoxon test at the 0.05 significance level. Please see section 5.5.3 for further details on why the specific comparison to model D is being made. Most importantly, RF and NN did not perform better in accuracy than LR statistically. That said, LR model performed better than SVM. Thus, given the advantages of LR for assessment of fit and interpretability, we chose this as the focus of further study and development.

The null hypothesis is that the accuracy of LR when compared to algorithms accuracy such as from RF, SVM, and NN is from the same population Higgins (2003). This can be described more formally as follows:

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

Please note that $A2$ in the comparison denotes RF, SVM, or NN. The results of the comparison and tests are given below:

Table 3: LR Model D Accuracy Comparison LOOCV

Comparison	P-Value
LR and RF	0.5841
LR and SVM	0.0129
LR and NN	0.9129

Table 4: LR Model D Accuracy Comparison 2FCV

Comparison	P-Value
LR and RF	0.3401
LR and SVM	0.0027
LR and NN	0.2973

Table 5: LR Model D Accuracy Comparison 3FCV

Comparison	P-Value
LR and RF	0.5287
LR and SVM	0.0003
LR and NN	0.6070

First, this paper rejects the null hypothesis of accuracy scores between LR and SVM for model D. The comparisons fail to reject the null hypothesis when LR was compared to RF and NN respectively. The results are similar when comparing other LR models to

other machine learning algorithms on accuracy and AUC. Therefore, since LR is statistically similar to NN and RF in accuracy and AUC performance, it follows this analysis would use of general linear model theory on LR for fit and interpretation of coefficients to build an explanatory model. The comparisons reveal that LR through model D is a viable candidate machine learning algorithm to detect and predict arousal in participants given biometric and built environment characteristics.

5.2. Logistic Model Fit

In this section, this paper will examine the model performance of LR across LOOCV, 2FCV, and 3FCV to assess model fit. Specifically, we will look at the accuracy, AUC, AIC, PR2, and Chi-Square results.

Table 6: LOOCV LR Accuracy, AUC, AIC, PR2, and Chi-Square

Algorithm	Model	Accuracy	AUC	AIC	PR2	Chi-Square
LR	A	0.5632	0.4977	14836.26	0.0211	0*
LR	B	0.7075	0.7239	11663.41	0.2354	0*
LR	C	0.6532	0.6643	7053.463	0.5391	0*
LR	D	0.7220	0.7357	11491.89	0.2470	0*
LR	E	0.7563	0.7842	10001.08	0.3418	0*

The $*$ means that every Chi-Square fit tests in the fold rejected the null hypothesis at 0.0000 meaning all models in the table above are statistically more useful than the null average intercept only model. From the above, model A had the worst accuracy, AUC, and fit. Model C, the full model, had the best fit score of 0.5391 but the lower accuracy score shows this model overfits the training data. Model B and Model D are comparable in accuracy and AUC scores. However, model D has a lower AIC score than model B. Alone, this would suffice in preferring model B given accuracy and AUC are comparable. In addition, model D also has a better pR^2 score and therefore fits the training data better. Model E by the metrics used above looks very competitive with accuracy, AUC, AIC, and fit values competitive with the other models. However, the next section will reveal why this model should likely not be chosen.

Table 7: 2FCV LR Accuracy, AUC, AIC, PR2, and Chi-Square

Algorithm	Model	Accuracy	AUC	AIC	PR2	Chi-Square
LR	A	0.5399	0.5005	13910.33	0.0233	0*
LR	B	0.7548	0.7567	11084.73	0.2270	0*
LR	C	0.5179	0.5449	5646.649	0.6154	0*
LR	D	0.7566	0.7585	10921.22	0.2387	0*
LR	E	0.8096	0.8266	9526.058	0.3329	0*

Again, model A continues its poor performance. The full model has excellent fit, but poorer accuracy and AUC which indicates overfit. Here, it appears that model B and D

have almost indistinguishable accuracy and AUC. However, the average AIC and pR^2 clearly favor model D over model B. Model E again has good accuracy, AUC, and fit metrics.

Table 8: 3FCV LR Accuracy, AUC, AIC, PR2, and Chi-Square

Algorithm	Model	Accuracy	AUC	AIC	PR2	Chi-Square
LR	A	0.5487	0.4983	13020.57	0.0277	0*
LR	B	0.7734	0.7641	10547.08	0.2180	0*
LR	C	0.5248	0.5537	4895.327	0.6395	0*
LR	D	0.7705	0.7606	10355.5	0.2326	0*
LR	E	0.8183	0.8292	9089.698	0.3231	0*

3FCV is the most general fit and therefore the most ideal validation considered in this paper. Thus, extra attention should be paid to the results above. The trends observed in the other folds continue. Model A has performed poorly. The full model, model C, has excellent fit but the accuracy and AUC reveal the overfit issue is persistent. Models B and D has comparable accuracy and AUC scores, but the average AIC and pR^2 clearly favor model D being a superior fit. Most importantly, the accuracy and AUC scores have gone up, which indicates that the models are not overfitting on the data as they were for LOOCV.

Given the above, we now reflect on the interpretation of the coefficients of the LR algorithm. Based on the accuracy and fit metrics discussed above, it has been shown that 3FCV LR model D is a viable model for detecting and predicting arousal in participants given biometric and built environmental characteristics and the criteria above. The explanatory implications of the model are further described below. This paper will also briefly discuss why model E should be discarded.

5.3. Explanatory Model

This paper looked at the coefficients for 3FCV for model D and E. Despite model E having competitive accuracy, AUC, and fit metric scores, it has been discarded because the coefficients in the model for Zone were not statistically significant at 0.05 level. Said more concretely, the addition of the zone variable increased the performance metrics above but did not statistically contribute to explaining arousal in the model. In a sense, model E was the most appropriate model from the experimental design perspective since the experiment relied upon experimental zones and participants. Therefore, it was no surprise that model E had good accuracy, AUC, and fit. However, from a explanatory model perspective, the zone does not adequately capture the environmental characteristics as other models, especially model D that has walkability, Number of Lights, Trees, Lines, Points, and Grass. In other words, while model E might be an interesting model from a machine learning centric approach, it fails as a good explanatory model. Therefor, 3FCV was considered as a template for an explanatory model. The folds and performance of 3FCV were very similar for all 6 folds. Consequently, it suffices for us to consider the results for fold 1. In addition, this section will only discuss the variables that are statistically significant in the model. See below for the explanation of the coefficients:

Table 9: 3FCV LR Model D Statistically Significant Coefficients

Coefficient	Estimate	Std. Error	P - Value
Norm HR	-0.96054	0.13778	3.14e-12
Norm EDA	0.52031	0.10942	1.98e-06
Walkability	-0.42300	0.03400	2e-16
Num Lights 6	0.55417	0.14641	0.000154
Num Lights 9	0.96066	0.24840	0.000110
Num Lights 17	0.62360	0.24744	0.011728
Num Trees 2	-0.57888	0.12320	2.62e-06
Num Trees 5	-0.49781	0.11358	1.17e-05
Num Trees 6	-0.73623	0.11477	1.41e-10
Num Trees 7	-0.97217	0.13441	4.73e-13
Num Trees 8	-1.32633	0.16632	1.53e-15
Num Trees 9	-1.48456	0.15702	2e-16
Num Tees 10	-2.94440	0.61184	1.49e-06
Num Lines	0.41176	0.02682	2e-16
Num Points	0.22438	0.07942	0.004728
Num Grass 1	0.62416	0.08694	7.00e-13
Num Grass 2	1.47801	0.10110	2e-16
Num Grass 3	2.86276	0.11734	2e-16
Num Grass 4	2.81924	0.15745	2e-16
Num Grass 5	2.52025	0.43669	7.87e-09
Num Grass 6	2.98681	0.54243	3.66e-08

These results are mostly significant at an 0.001 level and all at a significance level of 0.05. The coefficients were tested individually in assessing if their addition contributed to explaining the variation of the model in a meaningful way. For example, Norm HR is highly significant in contributing to explaining arousal in the model at a p-value of nearly 0. Before this paper proceeds further, let us briefly mention that the levels in number of lights were from 1 to 19, but only 6, 9, and 17 contributed to explaining arousal in the model. For example, if a participant observed lights other than 6, 9, or 17 then it follows that the terms for lights in the model would be 0 and as discussed in section 5.2.5, would have a neutral contribution to arousal. Similarly, the number of trees had levels 2 to 13 but only levels 2, 5, 6, 7, 8, 9, and 10 were statistically significant. Interestingly, all levels for grass were statistically significant.

The model suggests that higher heart rate and presence of trees contribute to the likelihood of the user to have low arousal and therefore some association with feeling safer. Conversely, the model suggests higher rates of EDA, lines, points, and grass contribute to the individual having a higher likelihood of experiencing an arousal event and therefore feeling less safe. While the findings may be useful for interpretation, they should be cautiously considered until further studies can be conducted. Nevertheless, the primary outcome demonstrates that the proposed machine learning methodology can be used to identify specific characteristics that cause arousal.

There are some important caveats to mention. The authors assert that this model is useful and explanatory of the data fitted by the LR machine learning algorithm. The model is not meant to be generalized beyond the current data and be inferred on the general population. The results, however, are suggestive, and future work should focus on such a task. In addition, the models have only focused on main effects and not interactions. Statistical interactions are likely to be highly informative both to machine learning and built environment researchers. Nevertheless, the results in this paper establish a connection between a participants biometrics, environmental variables, and the detection of their arousal affects via machine learning.

6. Summary

The authors assert that future work should focus on increasing the scale of the experiment, diversify and increase the number of built environments in the experiment, balance participants by gender, and further weaken potential cofounding influences in the experiment by requiring half of the participants to walk the route counter-clockwise and the other half clockwise.

Many machine learning researchers have noted that there is no single best model that works optimally for all kinds of problems Robert (2014). However, the results in this analysis certainly suggest that binary classification machine learning algorithms provides a useful approach and methodology in the detection of affect in a built environment. First, the Chi-Square tests reveal that the models are statistically useful in explaining affect over the null or intercept model. Second, AIC provide a useful measure in suggesting model D as an appropriate model over the others. Model E was competitive, but the Wald test revealed it was an inappropriate explanatory model. In summary, pR^2, Wald tests, and the accuracy and AUC models indicate that LR model D fits the data well, predictive abilities, and contains useful explanatory information about the data collected in the course of the experiment.

References

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.

Alan Agresti. *Logistic regression*. Wiley Online Library, 2002.

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.

M Bishop Christopher. *PATTERN RECOGNITION AND MACHINE LEARNING*. Springer-Verlag New York, 2016.

Casey Erdmier, Jason Hatcher, and Michael Lee. Wearable device implications in the healthcare industry. *Journal of medical engineering & technology*, 40(4):141–148, 2016.

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

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

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Maurizio Garbarino, Matteo Lai, Dan Bender, Rosalind W Picard, and Simone Tognetti. Empatica e3a wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on*, pages 39–42. IEEE, 2014.

Paul Gilbert, Kirsten McEwan, Ranjana Mitra, Leigh Franks, Anne Richter, and Hellen Rockliff. Feeling safe and content: A specific affect regulation system? relationship to depression, anxiety, stress, and self-criticism. *The Journal of Positive Psychology*, 3(3): 182–191, 2008.

David Giles, Nick Draper, and William Neil. Validity of the polar v800 heart rate monitor to measure rr intervals at rest. *European journal of applied physiology*, 116(3):563–571, 2016.

Jennifer Healey and Beth Logan. Wearable wellness monitoring using ecg and accelerometer data. In *Wearable Computers, 2005. Proceedings. Ninth IEEE International Symposium on*, pages 220–221. IEEE, 2005.

Jennifer A Healey and Rosalind W Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on intelligent transportation systems*, 6 (2):156–166, 2005.

David A Hensher and Peter R Stopher. *Behavioural travel modelling*. Taylor & Francis, 1979.

James J Higgins. Introduction to modern nonparametric statistics. 2003.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

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.

Saewon Kye, Junhyung Moon, Juneil Lee, Inho Choi, Dongmi Cheon, and Kyoungwoo Lee. Multimodal data collection framework for mental stress monitoring. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, pages 822–829. ACM, 2017.

Brett Lantz. *Machine learning with R*. Packt Publishing Ltd, 2015.

Florian Lederbogen, Peter Kirsch, Leila Haddad, Fabian Streit, Heike Tost, Philipp Schuch, Stefan Wüst, Jens C Pruessner, Marcella Rietschel, Michael Deuschle, et al. City living and urban upbringing affect neural social stress processing in humans. *Nature*, 474(7352): 498, 2011.

Cameron McCarthy, Nikhilesh Pradhan, Calum Redpath, and Andy Adler. Validation of the empatica e4 wristband. In *Student Conference (ISC), 2016 IEEE EMBS International*, pages 1–4. IEEE, 2016.

Daniel McFadden. The measurement of urban travel demand. *Journal of public economics*, 3(4):303–328, 1974.

Leslie Mertz. Convergence revolution comes to wearables: Multiple advances are taking biosensor networks to the next level in health care. *IEEE pulse*, 7(1):13–17, 2016.

David Metcalf, Sharlin TJ Milliard, Melinda Gomez, and Michael Schwartz. Wearables and the internet of things for health: Wearable, interconnected devices promise more efficient and comprehensive health care. *IEEE pulse*, 7(5):35–39, 2016.

David Meyer. Support vector machines: The interface to libsvm in package e1071. 2004.

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

John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.

R Parker. Your environment and you: Investigating stress triggers and characteristics of the built environment. 2016.

Brian Ripley, William Venables, and Maintainer Brian Ripley. Package nnet. *R package version*, pages 7–3, 2016.

Christian Robert. Machine learning, a probabilistic perspective, 2014.

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.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533, 1986.

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.

Lipo Wang. *Support vector machines: theory and applications*, volume 177. Springer Science & Business Media, 2005.

T Whitaker. Linking affect and the built environment using mobile sensors and geospatial analysis. 2018.

Heath Yates, Brent Chamberlain, and William H Hsu. A spatially explicit classification model for affective computing in built environments. In *Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), 2017 Seventh International Conference on*, pages 100–104. IEEE, 2017a.

Heath Yates, Brent Chamberlain, Greg Norman, and William H Hsu. Arousal detection for biometric data in built environments using machine learning. In *IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*, pages 58–72, 2017b.