

Affective Intelligence in Built Environments

by

Heath Landon Yates

B.S., University of Missouri - Kansas City, 2004

M.S., Emporia State University, 2008

M.S., Kansas State University, 2011

M.S., Kansas State University, 2014

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2018

Abstract

The contribution of the proposed dissertation is the application of affective intelligence in human-developed spaces where people live, work, and recreate daily, also known as built environments. Built environments have been known to influence and impact individual affective responses. The implications of built environments on human well-being and mental health necessitate the need to develop new metrics to measure and detect how humans respond subjectively in built environments. Detection of arousal in built environments given biometric data and environmental characteristics via a machine learning-centric approach provides a novel and new capability to measure human responses to built environments. Work was also conducted on experimental design methodologies for multiple sensor fusion and detection of affect in built environments. These contributions include exploring new methodologies in applying supervised machine learning algorithms, such as logistic regression, random forests, and artificial neural networks, in the detection of arousal in built environments. Results have shown a machine learning approach can not only be used to detect arousal in built environments but also for the construction of novel explanatory models of the data.

Affective Intelligence in Built Environments

by

Heath Landon Yates

B.S., University of Missouri - Kansas City, 2004

M.S., Emporia State University, 2008

M.S., Kansas State University, 2011

M.S., Kansas State University, 2014

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Computer Science
College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2018

Approved by:

Major Professor
William H. Hsu

Copyright

© Heath Landon Yates 2018.

Abstract

The contribution of the proposed dissertation is the application of affective intelligence in human-developed spaces where people live, work, and recreate daily, also known as built environments. Built environments have been known to influence and impact individual affective responses. The implications of built environments on human well-being and mental health necessitate the need to develop new metrics to measure and detect how humans respond subjectively in built environments. Detection of arousal in built environments given biometric data and environmental characteristics via a machine learning-centric approach provides a novel and new capability to measure human responses to built environments. Work was also conducted on experimental design methodologies for multiple sensor fusion and detection of affect in built environments. These contributions include exploring new methodologies in applying supervised machine learning algorithms, such as logistic regression, random forests, and artificial neural networks, in the detection of arousal in built environments. Results have shown a machine learning approach can not only be used to detect arousal in built environments but also for the construction of novel explanatory models of the data.

Table of Contents

List of Figures	x
List of Tables	xi
Acknowledgements	xi
Dedication	xii
Preface	xiii
1 Introduction	1
1.1 Introduction	1
1.2 Challenges	2
1.3 Objectives	3
1.4 Definitions	4
2 Affective Computing	6
2.1 Background	6
2.2 Affective Computing	7
2.2.1 The Beginning: Late 90s	7
2.2.2 The Early Years: 2000 - 2010	8
2.2.3 Rapid Advances and Commercialization: 2010 - 2018	10
2.3 Built Environments	11
3 Arousal Detection for Biometric Data in Built Environments using Machine Learning	13
3.1 Introduction	13

3.1.1	Goals	13
3.1.2	Limitations of Existing Work	14
3.1.3	Objectives and Significance	15
3.1.4	Central Hypothesis	15
3.1.5	Approach	16
3.2	Background and Related Work	18
3.2.1	Related Work	18
3.2.2	Heart Rate and Arousal Estimation	19
3.2.3	Support Vector Machines	19
3.2.4	Multilayer Perceptrons	20
3.2.5	Logistic and General Linear Mixed Models	20
3.3	Methodology	21
3.3.1	Data Preparation	21
3.4	Experiment Design: Evaluation Strategy	23
3.4.1	Model Selection and Discrimination Strategy	23
3.4.2	Cross-validation Strategy and Calibration Strategy	23
3.5	Experiment Design: Results	24
3.5.1	Normalized Heart Rate by Question Zone	24
3.5.2	Model Selection and Discrimination	26
3.5.3	Model Calibration	27
3.6	Summary	29
4	A Spatially Explicit Classification Model for Affective Computing in Built Environments	30
4.1	Introduction	30
4.2	Arousal Detection in Built Environment	31
4.2.1	Motivation for Arousal Detection in Built Environments	31
4.2.2	Detecting Arousal in Built Environment	32

4.2.3	Proposed Guidelines and Open Questions	35
4.3	Design for Arousal Detection in Built Environments	35
4.3.1	Data Provenance	36
4.3.2	Data Processing and Analysis	37
4.4	Summary	39
5	Binary Classification of Arousal in Built Environments using Machine Learning	40
5.1	Introduction	40
5.1.1	Goals	40
5.1.2	Motivations	41
5.1.3	Objectives and Significance	41
5.1.4	Central Hypothesis	42
5.2	Background and Related Work	43
5.2.1	Related Work and Wearables	43
5.2.2	Random Forests	44
5.2.3	Support Vector Machines	44
5.2.4	Multilayer Perceptrons	45
5.2.5	Logistic Regression	46
5.3	Experiment Design: Methodology	47
5.3.1	Approach	47
5.3.2	HR and EDA Normalization	49
5.3.3	Arousal Prediction Target	50
5.3.4	Data Preparation	51
5.4	Experiment Design: Evaluation Strategy	53
5.4.1	Model Selection and Discrimination Strategy	54
5.4.2	Cross-validation Strategy and Calibration Strategy	55
5.5	Experiment Design: Results	58
5.5.1	Accuracy and AUC Scores	58

5.5.2	Comparisons	62
5.5.3	Logistic Model Fit	64
5.5.4	Explanatory Model	66
5.6	Summary	69
6	Summary and Future Work	70
6.1	Summary	70
6.2	Conclusions	71
6.3	Limitations	72
6.3.1	Open Questions	72
6.4	Future Work	74
6.4.1	Short-Term	74
6.4.2	Long-Term	75
7	Code References	78
	Bibliography	80

List of Figures

3.1	Mean Normalized HR and 95% Confidence Interval by Question Zone	25
4.1	Map of Different Environments Encountered on Walk	33
4.2	Mean Normalized HR and 95% Confidence Interval by Different Environments Encountered	34
4.3	Arousal Detection using Wearables and Sensors in Affective Computing Context	38
5.1	Multilayer Perceptron	45
5.2	Urban Built Environment Annotated with Zones	47
5.3	Experimental Zone Z7	48
5.4	Designated Route for Walk by Each Participant	48
6.1	Continuous Zones	75

List of Tables

3.1	Schema of Data	22
5.1	Schema of Data	52
5.2	Model Specification	53
5.3	Leave One Out Cross-validation	56
5.4	2 Fold Cross-validation	57
5.5	3 Fold Cross-validation	57
5.6	LOOCV Accuracy and AUC	59
5.7	2FCV Accuracy and AUC	60
5.8	3FCV Accuracy and AUC	61
5.9	AN and AP Accuracy	62
5.10	LR Model D Accuracy Comparison LOOCV	63
5.11	LR Model D Accuracy Comparison 2FCV	63
5.12	LR Model D Accuracy Comparison 3FCV	63
5.13	LOOCV LR Accuracy, AUC, AIC, PR2, and Chi-Square	64
5.14	2FCV LR Accuracy, AUC, AIC, PR2, and Chi-Square	65
5.15	3FCV LR Accuracy, AUC, AIC, PR2, and Chi-Square	65
5.16	3FCV LR Model D Statistically Significant Coefficients	67

Acknowledgments

This dissertation research would not have been possible without the kind support and assistance from family, friends, and colleagues. Thank you to my advisor, Dr. William H. Hsu, for his mentorship and guidance throughout this process. Words alone do not convey my appreciation. In addition, I would like to extend my thanks to Dr. Brent Chamberlain for his friendship, assistance, and mentorship. Deep thanks to my PhD committee for their kind assistance and support.

I would also like to extend my profound thanks and appreciation to the Biosecurity Research Institute. Especially the generous assistance and support from Will Baldwin. I thank him for his wisdom, friendship, and advice as a professional mentor. Special thanks to Dr. Dana Vanlandingham, who in addition to serving on my PhD committee, also for guided and supported me at the institute as well. For their guidance, mentorship, and support, I also thank Dr. Weixing and Dr. Xiuqin Bai.

Thank you to Dr. Jennifer Healey for her kind mentorship, collaboration, and assistance. I deeply value her guidance on biometric data and how to deal with normalization issues in data. She has shown me the type of scientist I would like to become.

Of course, I must also thank my friends for welcoming me into the computer science family and accepting me as one of their own. I convey my profound thanks to Brian Cain, Kevin Dean, Chris Iszory, Russ Feldhausen, Kevin Radcliffe, Carlos Salazar, and Jayson Sharp. I would also like to especially thank Brian Sweeney and Joshua Weese for their friendship, guidance, and mentorship.

I thank my wife, children, and family for their unconditional love and support, without which none of this would have been possible. Thank you.

Dedication

I firstly dedicate this Ph.D. to my loving and beautiful wife, Chunfang Yates. I secondly also dedicate this work to my children, mother, father, and sister. Always in my heart, I lastly dedicate this work to my grandfather Herschell Gragg and my grandmother Kathryn Gragg.

Preface

This dissertation is comprised of seven chapters. The following chapters were written entirely by me: Chapter 1, Chapter 2, Chapter 5, Chapter 6, and Chapter 7. Please note that Chapter 7 describes source code developed by me during the course of this dissertation. It also links to a code repository created by me. The raw data belongs to Dr. Brent Chamberlain and all inquiries should be made to him for request.

A version of chapter 2 has been submitted for publication. Not included in that submission are the sections on built environments, emotions, and wearables.

A version of chapter 3 has been submitted for publication. I wrote the manuscript and am responsible for developing the arousal detection using wearables and sensors in affective computing context. Figure 3.1 and Figure 3.2 were developed by Dr. Brent Chamberlain and Parker Ruskamp. Dr. Brent Chamberlain, who is the second author on the paper, contributed in section 3.2.2 discussing the work he oversaw with his student Parker Ruskamp and some open questions in section 3.4.1. He also provided substantial review of the manuscript. Dr. William Hsu, who is last author, also provided a significant review of the manuscript.

A version of chapter 4 has been submitted for publication. I wrote the manuscript and am the sole author of the code, algorithm preparation, statistical analysis, model selection, calibration, and algorithms result summary. Dr. Brent Chamberlain is second author on the publication, provided insights on normalization of heart rate, and provided a substantial review of the manuscript. The raw data was collected by Parker Ruskamp under the guidance of Dr. Brent Chamberlain. Dr. Chamberlain also is responsible for the zones experimental design structure in the raw data. In addition, Dr. Brent Chamberlain and I worked together on a merged zone to biometric data set. Dr. Greg Norman provided a domain expert annotation of the raw data for arousal binary classification. Dr. Hsu, who is last author, also provided a significant review of the manuscript.

Chapter 5 was written completely by myself, except for figures 5.2, 5.3, and 5.4, which were developed by Taylor Whitaker. The raw data was collected by Taylor Whitaker and Parker Ruskamp under the guidance of Dr. Brent Chamberlain. The polar data joined with built environment variables was prepared by Taylor Whitaker under the guidance of Brent Chamberlain. In addition, Dr. Chamberlain helped instruct the experimental design of the raw data.

Chapter 6 was written completely by me. The first paragraph in section 6.3.1 presents interdisciplinary open questions that are inspired by the paper covered in chapter 4 with input from Dr. Brent Chamberlain. Figure 6.1 was developed by Dr. Brent Chamberlain for use in this chapter. The real time annotation sensor in future work will be a joint innovation between Will Baldwin and me.

Chapter 1

Introduction

1.1 Introduction

Affective computing is defined as the study and development of systems, computational devices, or wearables that can process, interpret, recognize, and detect human affects¹. This area of research is very active and is driven by several promising areas for applications such as virtual reality, smart surveillance, perceptual interfaces, etc². As early as 1960, Page advanced a control centric theory that proposed users someday would be able to share their thoughts in real time with machines³. Beyond this, there is an even wider long term picture in the pursuit of affective computing. Affective computing is also related to artificial intelligence. For example, the HAL 9000 computer in the Kubrick and Clarke film *2001: A Space Odyssey* is an affective computer. HAL could recognize individuals, read lips, understand the aesthetics of art sketches, and recognize the emotions expressed by his astronaut colleagues⁴. This example also is a cautionary tale of affective abilities we also might not want to give machines, such as the ability to lie and do physical harm to humans. Ultimately though, affective computing is not about giving machines emotions, but the ability to have emotional intelligence and learn about the affective state of humans. Picard proposed in 2001 that machine intelligence needs to include emotional intelligence and that affective computing is a way to provide this capability to machines by processing physiological signals to infer

information about human affects⁵. Recently, industry has been trying to do this as well. Affective computing research is progressing and finding new applications and use cases as time progresses.

The benefits outweigh the risks. For example, besides artificial intelligence, affective computing has been proposed as a way to combat drug addiction⁶. Researchers have also developed affective wearable devices that report on social-emotional information in real-time to assist individuals with autism⁷. This dissertation suggests another area that deserves attention is built environments. This dissertation contends research questions associated with built environments is a rich field with considerable potential concerning the application, refinement, and advancement of machine learning algorithms in built environments to detect human affects in a built environment. The interdisciplinary nature of this research holds promise to each researcher of varying background by the very merit of the multifaceted nature of the research problem.

Affective computing shows promise in teaching machines to be given the ability to sense and recognize expression of human emotions such as interest, distress, pleasure, and stress, with the goal of having a machine choose more helpful and less aggravating behaviour⁸. This dissertation proposes to show that affective computing can be applied to assess human affect in built environments and therefore stimulate future research and development into better design of built environments to enhance the human experience in said environments and tools to continually measure human affective responses to built environments.

1.2 Challenges

In order to use data accurately, data must be annotated appropriately⁹. It is challenging to detect accurately physiometric states of a user or their emotional state even in laboratory settings, thus it is of direct interest to account for confounding variables in a natural environment¹⁰. In addition, it has been shown by Picard and Healey that various types physical activities can cause changes in physiological output¹¹ and can also be a potential confounding factor. Therefore, our future relevant studies should attempt to instruct users to focus

on one type of physical activity, like walking, rather than mixing walking and running when interacting with a built environment.

Emotion recognition is difficult since emotion is not well defined¹². In addition, affective state information is hard to measure¹³. One of the challenges that has faced affective computing since the inception of the field is that emotion theory has not determined what physiological patterns accompany each emotion¹⁴. It is challenging to infer standard patterns on human affect states on a wide range of humans given individual pattern differences due to gender, personality, and ethnic background¹¹.

Affective computing is still a young and growing field, and one of the challenges acknowledged by researchers is how to bring together theorists and applied practitioners from different fields to refine nomenclature used with respect to affect and learning.¹³

1.3 Objectives

This Ph.D. dissertation proposes three novel contributions as follows:

1. Design methodology for affective computing in built environments

Literature review suggests that while some work has been tangentially related to questions associated with affective computing, environment, and built environments that the application of formal research in affective computing in built environments is new, nascent, and an open area of inquiry. Therefore, this necessitates the need to develop a design methodology for the detection of affect in a built environment.

2. Pertinent variables from built environment and sensors to detect user arousal in built environments

One of the most persistent problems in affective computing, no matter what the application domain is, is to determine what the appropriate features are from sensor information to detect the human affect signal of interest. Concretely, the detection of arousal or commonly called stress could be skin conductivity, average heart rate, average respiration rate, or a combination of these signals⁹. Classic results in literature suggest that galvanic skin response and heart rate may be correlated with affective states of an individual¹⁵⁻¹⁷. Healey

has argued that a multimodal approach is necessary. Most importantly, it will be useful for researchers to consider what might be the most important environmental variables and metrics to measure potential effects on human affects in built environments.

3. Applying, refining, and advancing the application of machine learning in affective computing for built environments

Picard has noted that emotion research will advance forward when it changes focus from summary and primitive statistical comparisons based on self reported data in a laboratory to characterizing patterns of data from individuals experiencing emotion in real life¹⁸. This dissertation proposes an approach to detect arousal in a built environment via a machine learning centric approach while using affective computing as the unifying framework. This dissertation proposes the development of a model that can fit, predict, and explain an affective response in a built environments context. In short, this dissertation asserts machine learning centric approach to built environments enables affective intelligence in built environments.

1.4 Definitions

The following are core terminology and definitions which will be used in this dissertation:

- **Affective Computing** - defined as the study and development of systems, computational devices, or wearables that can process, interpret, recognize, and detect human affects¹
- **Affective Intelligence** - The capability of an algorithm to recognize, discern, and label affect
- **Arousal** - A physiological response of sense organs being stimulated to a point of perception¹⁹.
- **Built Environments** - defined as human developed space that is comprised of where people live, work, and recreate on a daily basis²⁰.

Affective computing is the framework that enables the development of affective intelligence to be possible. Arousal is a complicated affect regulatory system that is often associated with feelings of safety and anxiety, and is crucial for motivating many feelings and behaviors²¹. Built environments basically consist of human developed space, and in that space, we find many environmental characteristics, such as grass, trees, power lines, roads, and buildings. The goal of this dissertation is to demonstrate the viability of affective intelligence in built environments via a machine learning approach by relying on the terms as understood and defined above.

Chapter 2

Affective Computing

2.1 Background

The origins of affective computing were in the 1960s when researchers began to ponder the ability of humans to transfer their entire thoughts to machines and vice versa. This was at first referred to as “man-machine coupling”³. The initial approach relied on control theory, not programming. The initial enthusiasm was replaced by the realization that the dream of man-machine coupling was a difficult problem. Rosalind Picard and other researchers in the late 1990s shifted focus instead to enable computers the ability sense human emotional states²². Concretely, research became interested in human computer interaction and their applications in reducing user frustration, enabling communication of user emotion, tools to enable social emotional skills, and developing infrastructure and applications to handle affective information²³. One of the ultimate goals of affective computing is affective interactions, when emotional information is communicated by the user in a natural and comfortable way that is recognized by the computer and used to improve its interaction with the user²⁴.

This “Human-Computer Coupling” is known today as affective computing. Since the inception of affective computing, the field has focused on how to teach machines to interpret signals, such as facial expression, vocal intonation, muscular movement, gesture, respiration, and automatic nervous system²². Central to this approach is the coupling of humans and

computers, which means the parallel development of wearable sensors and computers is essential to the affective computing research endeavor. This has been a natural development. For example, affective states, such as depression and anger, have been shown to have adverse affects on human health by weakening the immune system and rendering an individual susceptible to viral infection^{1;25}. Therefore, researchers have relied on wearables in order to monitor stress and other physiometric phenomenon outside of a laboratory or clinical environment¹. Classic results in literature suggest that galvanic skin response and heart rate may be a correlated with affective states of an individual²⁶.

Researchers argue that embedding empathy into the design of interactive health systems can be potentially vital to the acceptance and success of affective computing devices in that context²⁷. In addition, studies have shown that designing computer interfaces with relational skills, such as empathy, social dialogue, nonverbal immediacy behaviors, and other behaviours led to the increased proactive viewing of health information²⁸.

Affective computing is also interested in exploring alternatives in asking directly how users feel and trying to infer this feeling from physiological signals instead²⁹.

2.2 Affective Computing

2.2.1 The Beginning: Late 90s

Affective computing research started picking up steam in the late 1990s. In 1997, Picard and Healey demonstrated the use of wearables to detect human affect states using respiration, galvanic skin response (GSR), blood volume pressure (BVP), and electromyogram (EMG, and heart rate (HR)¹. They also outlined several of the general challenges that have persisted in the field ever since wearables were used to infer human affect states which are discussed in section 1.2. A year later, they demonstrated that the same signals could be used to detect human affect states using pattern recognition via Fischer linear discriminant. They also considered the participants' affective state changes over an average period of three minutes when previous studies were typically one to ten second reactions³⁰. Researchers have also used

frustrating users on purpose in order to detect the users' emotional state¹⁴. Using galvanic skin conductivity, capturing a users attention with video, and using a wearable computer to detect stress in real time was pioneered by Healey with the StartleCam³¹. It demonstrated that you could detect when a human was startled using pattern recognition that relied on a time-reversed filter and convolution sum. That is, the signal from the conductivity skin sensor was smoothed but preserving signal changes due to being startled. Interesting, wearable jewelry, shoes, glasses, and other affective wearable devices were anticipated in the late 1990s by researchers as well¹². Some examples of this nascent research were glasses equipped with embedded sensors were used to detect confusion and interest from users with moderately high accuracy³². User annotated approaches to studying affective computing were also explored in the offline and online recognition of emotion from physiological data. Specifically, Picard had a trained actress, equipped with various sensors that measure EMG, BVP, and GSR, display the affects: neutral, anger, hate, grief, platonic, romantic, joy, and reverence, to be recorded by the computer. Sequential floating forward search (SFFS) and Fisher projection (FP) were applied that achieved a recognition rate of 50% to 80% depending if the algorithm was run offline or online³³. Healey studied driver stress by measuring a users face with a camera that recorded there facial expressions approximately once a minute while driving. In addition, she measured BVP, EKG, and GSR. The initial studied determined there were many confounding factors and participants signals would differ to similar stimuli differently over the period of the study³⁴.

2.2.2 The Early Years: 2000 - 2010

At the beginning of the early 2000s, research began to expand into other areas and grow more sophisticated. For example, affective medicine, which teaches computers the recognition of emotion using variables that respond to people with active listening, empathy, and sympathy³⁵. Work also continued on recognizing the affective state of a human while driving. Healey was able to recognize patterns at a rate of about 80% for human stress in drivers by using a linear discriminant function while also recording a participants respiration, EMG,

HR, and GSR³⁶. A interesting head nod and head shake detector was created by researchers who used hidden markov models that tracked pupils, pupil position, and directions of head movements to detect when a head nod or shake occurs³⁷. In 2002, researchers also started to look into using affective computing to assist individuals diagnosed with autism³⁸. Affective computing has also been used in monitoring stress and heart health in real time with a phone and wearable computer as early as 2002³⁹. Real-time facial feature tracking continued to evolve by tracking eyes and eyebrows in real time extracting features using PCA⁴⁰. Modeling also was becoming more sophisticated, where researchers used dynamic Bayesian networks (DBN) and mixture of hidden Markov models to classify drivers speech under stress⁴¹. By 2004, there was a study by DARPA to study emotional state recognition to determine potential criminal intent which also sparked debate about ethics in affective computing⁴². Researchers also proposed around that time that affective systems should be minimalist in ubiquitous interface design. That is, the combination of parsimony and transparency where parsimony means a user interface is simple as possible and transparency by being designed to reduce cognitive demands as much as possible⁴³. It has also been shown that when agents are designed to simulate care and empathy, individuals who perceive the agents as caring are willing to continue working with them⁴⁴.

By 2005, the situation continued to evolve and find many new applications. Research has also been conducted into considering affective computing in adversarial situations, such as poker or interviewing experiments⁴⁵. One interesting result was showing how good typography can induce a good mood, which in turn has been shown to induce individuals to perform better cognitively and more creatively⁴⁶. Examination of stress also continued to advance, and grew in sophistication. Healey showed that you can use physiological signals to provide a metric of stress for individuals driving a vehicle and collect data for approximately 50 minutes.⁴⁷ Researchers used Bluetooth and a custom sensor called HandWave to measure skin conductance and used this to detect the emotional, cognitive, and physical arousal of users⁴⁸. Affective computing also was shown in 2005 to have potential in both educational learning environments and sensor fusion. Kapoor and Picard demonstrated that multimodal mixture gaussian models could achieve accuracy of over 86% in detecting if children were

interested or disinterested in solving a puzzle⁴⁹. A few years later, Kapoor went further and demonstrated a similar approach using gaussian and bayesian methods with 79% accuracy in detecting user frustration when using an intelligent tutoring application⁵⁰. Research also continued to look at the benefits of affective computing in assisting individuals who have autism⁵¹. Sensor technology itself started to advance in useful ways. In 2009, Poh and Picard developed a wristband for ambulatory assessment of electrodermal activity (EDA)⁵². This technology would eventually find its way into products on the commercial market like those offered by Empatica. Heart rate variability and electrodermal activity to classify children who had sensory impairment versus those who did not using sensors was explored using machine learning such as k-nearest neighbor (k-NN), decision trees, support vector machines (SVM), and linear discriminants with some accuracy⁵³. This demonstrated the potential viability of applying machine learning to affective computing problems.

2.2.3 Rapid Advances and Commercialization: 2010 - 2018

Affective Computing since 2010 has observed rapid advances in sensor technology, applications of machine learning, and commercialization of wearables. In 2010, Poh and Picard continued to perfect their wrist worn electrodermal activity (EDA) wrist band and showed it had advantages over FDA approved products at the time⁵⁴. Eydgahi and Picard proposed long term continuous monitoring of physiological data and the potential for commercial products to fit this niche⁵⁵. In the same year Fletcher, and Picard also proposed wearable sensors being used with mobile phones as a platform for advances in affective computing and low cost health care⁵⁶.

In 2011, researchers pioneered affective computing furniture in the way of a medical mirror that provided health information related to a person by tracking their daily vitals to aid in their health management⁵⁷. Researchers also trained computers to classify polite smiles versus actual smiles, the context was a business scenario between a banker and a client⁵⁸. Cardiovascular monitoring using earphones and a mobile device was also developed, which relied on a digital signal controller to process signals from sensors on the headphones⁵⁹

In 2013, a novel mirror sensor system and interface gave user feedback on their inner physiological responses with the goal of improving scientific understanding of psychophysiology in natural settings⁶⁰. One of the first applications of affective computing in political science was when researchers measured individuals candidate preferences based on video clips measuring their spontaneous facial reactions with accuracy over 74%⁶¹. Using mobile phones, researchers used the data collected from a mobile phone and wrist sensor over 5 days to accurately detect stress using binary classification of 75% based on surveys as a baseline⁶². In 2014, development in commercial wearables with high precision biosensors began to take off. Picard introduced the Empatica E3 wearable, a commercial product that has photoplethysmograph (PPG), electrodermal activity (EDA), 3-axis accelerometer, and temperature that uses bluetooth⁶³. Currently, this product has been superseded by the Empatica E4. It has been demonstrated that it is possible to measure cognitive stress remotely via heart rate variability using Naive Bayes and SVM as classifiers⁶⁴.

In 2015, Rivera demonstrated that wearables could be used toward the management of stress and detection by using Naive Bayes and SVM⁶⁵. Recent research also has begun exploring the potential of affective wearables to detect seizures in outpatient settings⁶⁶. The contemporary pace shows signs that the pace of advancements is accelerating and all evidence points towards more powerful wearable devices and sophisticated machine learning approaches, such as deep learning⁶⁷.

2.3 Built Environments

Research has shown there is a relationship between stress and a built environment^{68;69}. In addition, research has shown that stressful situations can induce an arousal response⁷⁰. This is discussed in further detail in section 3.1.5. However, current focus on machine learning and built environments is very little. The current state of literature in the field suggests a focus in using machine learning to advance capabilities in architecture, structures, and digital fabrication^{71;72}. Other approaches are using machine learning to analyze built environments from an economic perspective, such as property transactions for residential real estate⁷³. In

other words, the focus seems to be on how to leverage machine learning as a tool in the design of built environments or economic activities of built environments. Focusing on how built environments influence human affect via machine learning and implications in infusing built environments with affective intelligence could turn current research focus of machine learning and built environments on its head.

Chapter 3

Arousal Detection for Biometric Data in Built Environments using Machine Learning

3.1 Introduction

3.1.1 Goals

In this chapter, an approach is proposed 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. This chapter addresses 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 this chapter develops 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.

3.1.2 Limitations of Existing Work

Traditionally, it has been necessary to measure physiometric arousal indicators like 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 like walking or running on a treadmill while data are being recorded. The advent of wearables such as Empatica E4, Polar, Fitbit, and Garmin Vivosmart 3, provides researchers with the ability to conduct new experiments like measuring physiometric arousal indicators induced by a subject’s urban environment. These sensors include high quality pedometers, optical heart rate monitor, accelerometers, barometer, GPS, and GSR, allowing researchers 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. This chapter will 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 subject’s 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.

3.1.3 Objectives and Significance

This dissertation and chapter asserts that the following experimental approach will show promise in the classification and detection of arousal. By doing so, this dissertation will motivate the construction of a machine learning framework for developing a predictive intelligent system in future work. In addition, the work in this chapter will motivate the work in chapter 5 to determine which machine learning classification approaches are more appropriate over others.

This chapter 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 since the application domain field relies on expert annotation of arousal, developing an effective model for classification and prediction will demonstrate the viability this dissertation’s approach and provide a baseline for further research into intelligent systems. The third is to motivate future work which will both be covered further in chapters 5 and 6.

This chapter asserts that arousal detection using biometric, environmental variables, and neurophysiologist 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 chapter 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.

3.1.4 Central Hypothesis

This chapter presents two hypothesis 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. In this chapter, this dissertation will train models and then compare the mean accuracy and AUC of the tests. The conservative Mann-Whitney test which makes no parametric distributional assumptions is used. This chapter formalizes 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.

3.1.5 Approach

Evidence suggests linkages between a physical environment and its influence on mental health, well-being and human health^{68;74-77}. 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 mental⁷⁸⁻⁸⁰ 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^{81;82}. Unfortunately, the replacement of nature with built infrastructure may negatively impact public health at-large⁸³, leading to a greater risk of suffering from conditions like stress and mental fatigue^{78;80;84-86}.

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, researchers are now able to investigate relationships between a built-environment and human affective response in order to ascertain how design and planning of urban spaces may influence well-being. Research conducted by⁸⁷ and⁸⁸ suggests that there are strong physiological responses (e.g. reductions

in heart rate) to observing nature. Whereas their research was conducted in laboratory settings with discrete or short-term data, new machine learning techniques and wearable sensors⁸⁹ 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 relied on the Empatica E4 wearable and consisted of 12 subjects who volunteered to participate. The experiment was conducted by placing an Empatica E4 sensor on each individual subject. This sensor measured temperature, galvanic skin response, heart rate, time, and geospatial position.

Each individual walked along 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, a poorly lit street, well lit sidewalk, and calming park areas. For the experiment's control, each 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 have been cleaned and processed and are organized by participant ID. The data were processed and now have fields such like 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 have been trained on several applied machine learning techniques, such as SVM, ANN, LR, and GLMM. This chapter uses the standard methods provided by the R statistical language package for the comparison of ANN, SVM, and binary logistic regression methods. For example, ANN 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. Performance is measured by using accuracy and AUC.

3.2 Background and Related Work

3.2.1 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, like linear regression and support vector machines (SVM) to classify arousal⁹⁰. 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 is statistical in nature and presents an opportunity to build upon it using more sophisticated machine learning approaches. Users with higher levels of arousal measured reported appreciating the intervention feedback⁹¹. Sano has used binary classification with correlation analysis to determine physiological or behavioral markers for arousal using wearables and mobile computing⁹². The study showed that higher levels of arousal were correlated with activity level and screen on/off patterns⁹². Feasibility studies have shown that it is possible to classify and predict panic attacks using wearables and mobile computing using intelligent systems⁹³. The United States Army is showing interest in using wearables to collect real time information from soldiers⁹⁴. Future studies will likely consider arousal detection and prediction as well. Very recently, industry has been developing products that claim to detect stress in users as well by using variable heart rate⁹⁵.

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 results⁹⁶. Clearly, there is a lot of work that remains to be done in general

and in built environments.

3.2.2 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 safety⁹⁷. 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 walk⁹⁸. In an attempt to predict the general affective state of participants as 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 as “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 plus standard deviation change from the baseline condition for each participant.

3.2.3 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 *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 optimum⁹⁹. The SVM is a model that maps all observations upon a plane and divides them with a linear separator and margin. The linear separator and margin separates the observations into two classes. In other words, the decision boundary is chosen to be the one for which the margin is maximized¹⁰⁰.

3.2.4 Multilayer Perceptrons

Multilayer perceptrons (MLPs) are a type of feedforward artificial neural network (ANN) that, 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.

3.2.5 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 that relies on a logistic sigmoid acting on a linear function^{99;101}.

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, it is common to apply the data to a 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 function¹⁰². 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 are binary and also have autoregressive features.

3.3 Methodology

3.3.1 Data Preparation

The data have been cleaned and prepared. In addition, data are 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 table below 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, variables have been 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 square feet, and finally the variable to be estimated which is the presence of arousal or not. The data are shown in a table on the next page.

Table 3.1: *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

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.

3.4 Experiment Design: Evaluation Strategy

This section discusses variables present in the building of custom and novel machine learning models in this chapter. 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 are considered. In addition, the quality of EDA was not certain and ignored.

3.4.1 Model Selection and Discrimination Strategy

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

The analysis began simply with a full model and observed which variables were statistically significant. The analysis then proceeded further by removing, 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 model 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. This chapter discusses the results of this strategy in section 5.2.

3.4.2 Cross-validation Strategy and Calibration Strategy

In order to assess the accuracy of detecting arousal, the data are divided into training and testing data sets. Specifically, leave one out, 3 fold, and 4 fold cross-validation has been conducted. This dissertation chose four and three fold cross-validation only taking into account the small sample size and wanting to emphasize the possibilities, which in turn will be used to inform future studies. Since our data is longitudinal and organized by participant ID, the analysis proceeded by dividing the data into sections based on participant ID. For

example, for the first fold of 4 fold cross-validation the model was trained on participants 1, 5, 6, 7, 8, 9, 10, and 11. The analysis then accessed accuracy on participants 13, 14, 15, and 17.

Figure 1: Leave 4 Out 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 are divided into training and testing data sets by participant ID using 3-fold cross-validation as follows:

Figure 2: Leave 3 Out Cross-validation

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

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 it is believed worth merit to investigate model accuracy and AUC to inform subsequent studies.

3.5 Experiment Design: Results

3.5.1 Normalized Heart Rate by Question Zone

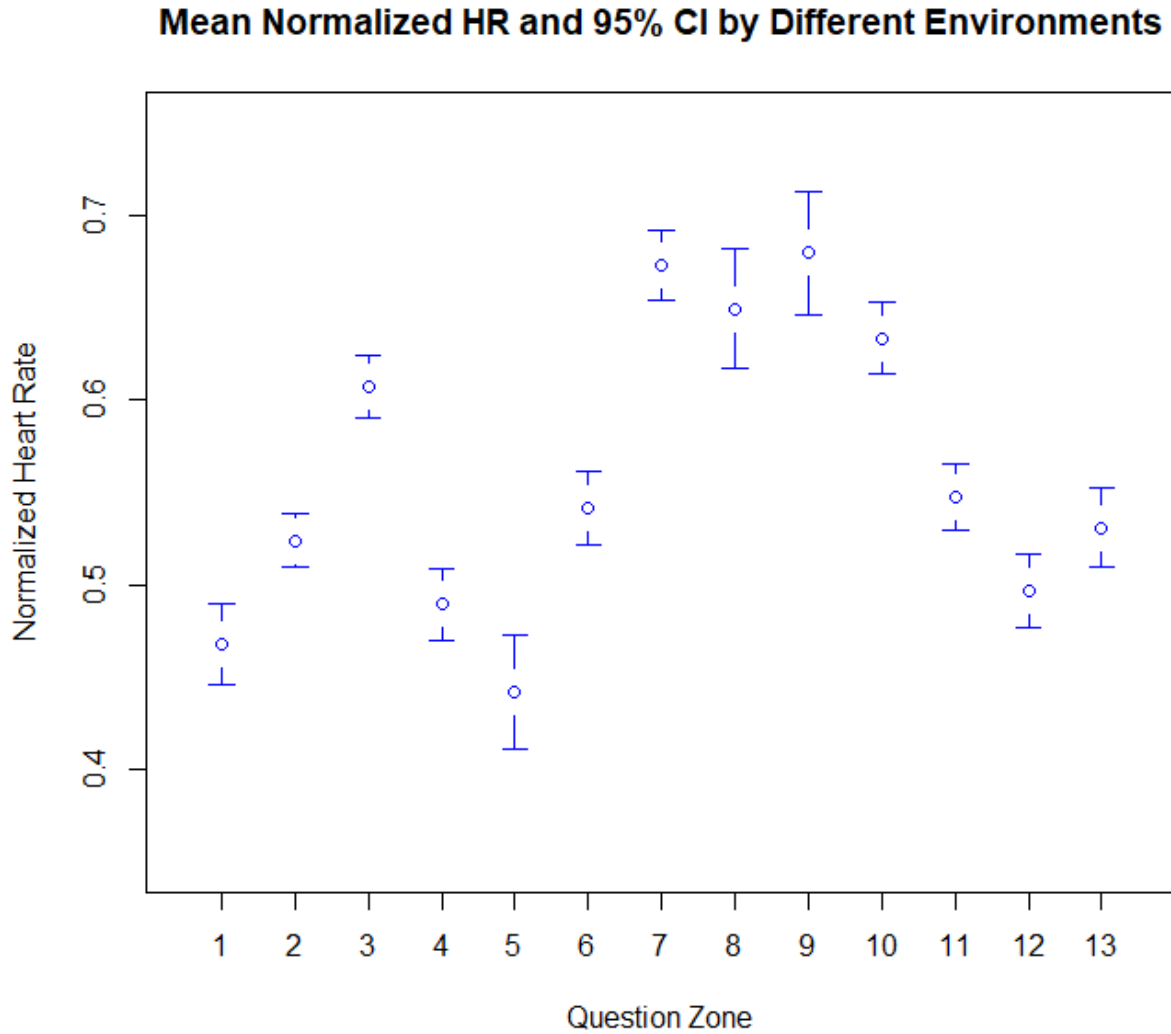
This chapter will examine the raw heart rate data of participants as varied by the question zones to see if there were any notable differences. The data were normalized as follows:

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

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. In this chapter, the analysis used the normalized heart rate and calculated the mean for all participants for each zone given the 95% confidence interval.

fig:NormMean

Figure 3.1: *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.

3.5.2 Model Selection and Discrimination

The scope of this chapter’s hypothesis, given the small sample size, was to assess the ability of machine learning models to fit the data in this context. This chapter fitted full models for SVM, ANN, logistic, and GLMM models. The approach was standard, removing one variable at a time in model specification until it came down to heart rate, temperature, and speed. The misclassification errors for all algorithms did not change that greatly, therefore opted to keep speed in the model given the preliminary nature of the analysis. This chapter assesses 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. The 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 are given below:

Figure 4: Misclassification Error Rate				
Algo	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 are binary in nature, future studies should focus both on smoothing over smaller windows (5 second, 10, and 15 second intervals instead of 30 seconds in this study) and collecting 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. This is discussed further in chapter 6 of this dissertation.

3.5.3 Model Calibration

This section will compare SVM, ANN, LR, and GLMM. Specifically, this section will compare accuracy and AUC. The conservative Mann-Whitney test revealed no statistical differences between the algorithms compared in accuracy and AUC. Consequently, this analysis 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: Leave 4 Out CV Accuracy			
Algo	Fold 1	Fold 2	Fold3
SVM	0.753	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. The AUC scores are as follows:

Figure 6: Leave 4 Out CV AUC			
Algo	Fold 1	Fold 2	Fold 3
SVM	0.474	0.591	0.731
ANN	0.193	0.825	0.675
LR	0.834	0.909	0.654
GLMM	0.836	0.903	0.655

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. This chapter compares SVM, ANN, and GLMM accuracy and AUC results below. Please note columns 1, 2, 3, and 4 refer to folds.

Figure 7: Leave 3 Out 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: Leave 3 Out 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

In regards to the error rate, since this is an exploratory study this dissertation considers an error rate of 10% or below to be tolerable. This chapter recommends more participants in

future studies. Thus, it is an open problem how to best reduce error rates in future studies. In general, for purely exploratory purposes of this dissertation, it will consider 75% or above an acceptable accuracy rate. By this metric, the results demonstrate the viability of machine learning given that all algorithms for the most part exceed the metric considered. That is, the methods examined here in context of built environment experiments are worth future study. Please see chapter 5 for further development.

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 size, but it is believed that further study is merited with more environmental variables should be included in subsequent studies.

3.6 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. In this chapter, this dissertation has shown the potential for these algorithms to have tolerable misclassification errors and demonstrated the potential for prediction by promising AUC scores. It is important to note the sample size is small and that statistical inference cannot be made. In addition, the results in this chapter did not differentiate which methods might be better. This is explored further in chapter 5 of this dissertation. Nevertheless, this chapter asserts that the potential for machine learning to be applied in this problem domain has been sufficiently suggested and merits further research.

Chapter 4

A Spatially Explicit Classification Model for Affective Computing in Built Environments

4.1 Introduction

Generally speaking, affective computing is the study and development of systems and devices that can recognize, interpret, and process human emotions or affects¹. Built environments are human-created settings that provide the infrastructure for human activities. These environments have been commonly defined as the human-made space in which people live, work, and recreate on a day-to-day basis²⁰. Architects and planners focus on creating built environments using a variety of forms and features that are designed to support society through functional approaches and human-centered use^{103;104}. Recently, there is increasing interest in exploring the association between built environments and mental health and other affects¹⁰⁵. This chapter asserts that an affective computing approach to questions on how built environments influence human emotions or affects is extremely relevant.

Through this chapter, this dissertation will also consider the motivation for using wearables and sensors as tools in studying human emotions given built environments. The asso-

ciation between affective computing and wearables goes back to the earliest days of research in affective computing⁹. In addition, there have been affective computing studies focused on detecting stress in participants¹⁰⁶. Recent advances in wearables and sensors are suggestive of how researchers might collect data in order to answer questions on how built environments influence human emotions. First, there is a general trend toward the growing availability, capabilities, and lowering costs of computing resources to process large quantities of data. Second, there is the growing sophistication and abilities of wearables and Internet of Things (IoT) devices. For example, sophisticated wearables such as the Empatica E4 can be used to collect a plethora of detailed physiological data in a non-entusive manner¹⁰⁷. In addition, commercial wearables such as Garmin’s vivosmart have purported abilities to detect stress in a user¹⁰⁸. Therefore, using existing and new capabilities of wearables and sensors in examining human affects in built environments is worth serious consideration.

This dissertation asserts that research shows persistent interest in how built environments can impact human affects and how wearables and sensors can be used to detect human affect within such environments. Therefore, this chapter shall focus on the following points. First, the uses and applications of affective computing systems associated with built environments focus on arousal detection. Second, it will briefly examine an experiment that will motivate guidelines for future work that is unobtrusive, effective, and likely to be more effectively used by participants in research. Third, as a guideline for affective and intelligent systems in a built environment context, this chapter will propose an approach and design for geospatial zone classification.

4.2 Arousal Detection in Built Environment

4.2.1 Motivation for Arousal Detection in Built Environments

Arousal can be defined as an elevated or different physiological state different from the average base physiological state like average heart rate¹⁰⁹. Measuring other physiological phenomenon, such as emotions and human affect, is difficult and complicated. In addition,

it is an open problem in correlating emotions to a physical space due to the complexity of the problem. This chapter acknowledges these challenges and believes framing the problem via the detection of arousal is an important first step in detecting human affect in built environments using wearable and associated sensors.

The detection of arousal using wearables and sensors will be useful in gaining insights into understanding built factors as an influence on human affects, and providing researchers with more sophisticated tools to test different strategies to alter emotion within a built environment[?]. Recent advances in wearable sensors¹¹⁰ and data analytics mean researchers are now able to collect new kinds of data in order to measure the relationship between built environments and human affective responses.

This chapter proposes that the detection of arousal with the use of wearables and sensors can help researchers investigate several issues of contemporary interest to landscape architects, planners and engineers. First, it is well known that perceived or actual unsafe built environments can adversely affect human mental health^{111–113}. In addition, there are risks from urbanization worth exploring in this context. Studies suggest that there is a correlation between a physical environment and human physical and mental health^{68;74–77}. There are also reasons to believe that natural elements in a built environment improve mental and physical health^{78–80}. Research by Ulrich and Parson et. al suggests that exposure to nature contrasted with built environments influences human affect and behaviors^{78;79}. The decline of nature globally due to rapid urbanization suggests a decline in public health⁸³, and consequently, this has the potential to increase the risks for adverse conditions, such as stress and mental fatigue^{78;80;84–86}. Therefore, an affective computing approach to the detection of human arousal in built environments should be the focus in collecting data to address these various concerns.

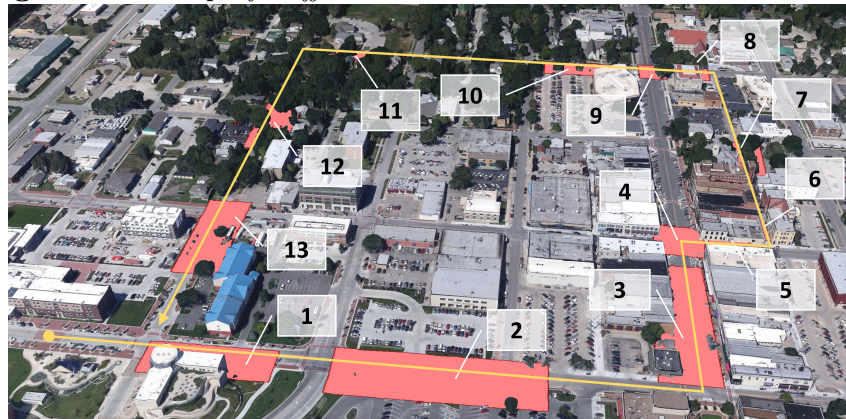
4.2.2 Detecting Arousal in Built Environment

Before this chapter proceeds further, it will briefly consider the findings of the previous chapter and its contribution in adding to the literature in the detection of arousal in built

environments, primarily using geospatial affective computing techniques¹¹⁴. This will provide further motivation for a spatially explicit classification model for affective computing in built environments.

Up to this point, preliminary work has focused on measuring how different environments encountered on a walk can be associated with physiological changes and differences. Specifically, Ruskamp⁶⁹ examined how different environmental characteristics affected arousal responses in participants in Manhattan, Kansas. The study had 12 college age students each fitted with an Empatica E4 and Polar sensor. Each participant was asked to walk a predetermined route that was chosen for specific environmental characteristics, such as a darkened alley, a poorly lit street, a well lit sidewalk, and areas with more vegetation features present like trees. These environmental characteristics were denoted by zones. Please see the map below regarding different environments encountered on the walk.

Figure 4.1: *Map of Different Environments Encountered on Walk*



After participants individually walked the route, each participant was given a survey in order to rate perceived safety of each zone on a Likert scale. The data outside of zones in the survey were not rated by participants. The responses in the survey were reported to be statistically significant by Ruskamp and Chamberlain⁶⁹, which reinforced the connection in literature that built environments are associated with human affect.

Using Ruskamp's data, last chapter examined the same data and experiment through an affective computing context. First, normalized heart rate collected from participants was examined and the mean for all participants was calculated for each question zone given

the 95% confidence interval. Each question zone corresponds to a specific environment of interest.

Figure 4.2: *Mean Normalized HR and 95% Confidence Interval by Different Environments Encountered*

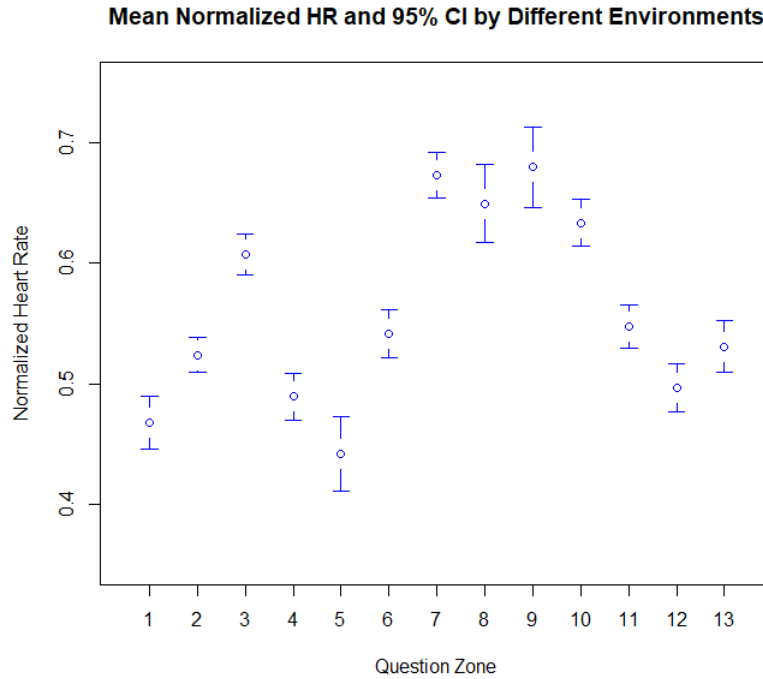


Figure 2 demonstrates statistically significant differences between the normalized HR by zone. Given the above, this is suggestive that different environments can have an influence on human physiological responses. Consequently, the last chapter looked at the control portion of the data to gain a baseline heart rate for each participant to compare against the data collected from the participant when they were walking the route. For the control, the user was asked to calmly walk from a predetermined starting point at a hotel as the beginning of the route. This 2 minute data collected on each participant were used to calculate the baseline for psychometric signals such as average heart rate. For each participant, the data from walking the route was smoothed using the mean heart rate and standard deviation grouped by every 30 seconds. These data were given to a neuropsychologist who looked at the smoothed and post survey data and gave an expert annotation of whether or not a participant was experiencing an arousal event for each zone. The last chapter applied

standard classification machine learning algorithms to the expert annotated data, and the results suggest that machine learning can be used to detect arousal in built environments using annotated data.

4.2.3 Proposed Guidelines and Open Questions

For studies applying affective computing to built environments, this chapter asserts that research should heavily consider adopting a geospatial approach toward data collection, which allows scientists to analyze data obtained from wearables and sensors in association with elements one experiences in space. It is very desirable to conduct similar experiments in as many different built environments as possible. In addition, seasonality, time of day, weather, and other natural environmental factors should be considered. Collecting such data over a long term duration may provide an opportunity to better estimate baseline data, identify significant events, and increase the reliability of data models. The present approach used expert defined user annotated data. This chapter will also motivate open questions for future experiments considering additional sensors that allow participants to directly annotate their responses in real time, not just post experiment in a questionnaire survey. This chapter asserts experiments using wearables and sensors are demonstrably unobtrusive, effective, and likely to be used by participants in experiments with considerable ease.

4.3 Design for Arousal Detection in Built Environments

As a guideline for affective and intelligent systems in a built environment context, this chapter proposes an approach and design for geospatial zone classification. Built environments are not found in laboratory settings. Consequently, measuring how differing environment spaces influence human affects requires an approach that allows researchers to measure physiological data in that space. Wearables and sensors in an affective computing context allows researchers in built environments a viable approach that can be used to detect a plethora of physiological phenomenon. Studies have demonstrated that wearables can be used in the

detection of stress⁹⁰. In addition, physiological data, such as heart rate, heart rate variability, electrodermal activity (EDA), and facial expressions have been used in classifying human affect^{58;106;110;115}. We propose that experiments conducted in the built environment domain space utilize wearables in order to collect data and use classification methods to detect human affect.

4.3.1 Data Provenance

This chapter will now briefly consider the nature of data that have been or could be collected by wearables in a built environment. For illustrative purposes, consider again, the study by Ruskamp⁶⁹ which relied on an Empatica E4 sensor. The sensor collected data like time, heart rate, temperature, and EDA. The study also collected additional data from participants through a survey asking them to rate different environmental zones. In addition, it is possible to process and annotate data by an expert for the presence of arousal or not. In the future, it might be possible to measure participant responses in real time through input devices, such as a mobile phone or another sensor to directly annotate the levels of arousal they are experiencing as they walk through different environments. Therefore, data provenance from expert annotated data is likely to be different from user annotated data. In other words, the expert annotated or user annotated scenario represents a different kind of data provenance that shall now be briefly described.

For expert defined annotation data of arousal, it follows that it is likely to be discrete data. That is, ordinal or binary data. Conversely, participant annotated arousal data could be continuous data that vary according to some scale, which are also described as nominal data. In both scenarios, arousal is a classification target will determine what sort of processing and analysis that will be appropriate¹⁰⁰. This chapter will now briefly consider how these data are likely to be designed.

Ordinal and Binary Data

Ordinal and binary data are likely to represent expert defined annotation of arousal. Specifically, the ground truth values that researchers elicit from subject matter experts that are holistic, subjective assessments of environmental arousal-induction level per geospatial zone. These are (1-7) Likert scale values elicited using annotation or survey questions and corresponding to discrete classification targets, namely, ordinal values. Similarly, expert annotation of arousal can take on a value of either 0 or 1 for each row of observation, where 0 indicates the absence of arousal and 1 the presence. Concretely, the presence or absence of arousal as defined by expert annotation is binary, and survey question results are discrete categorical variables¹¹⁶.

Nominal Data

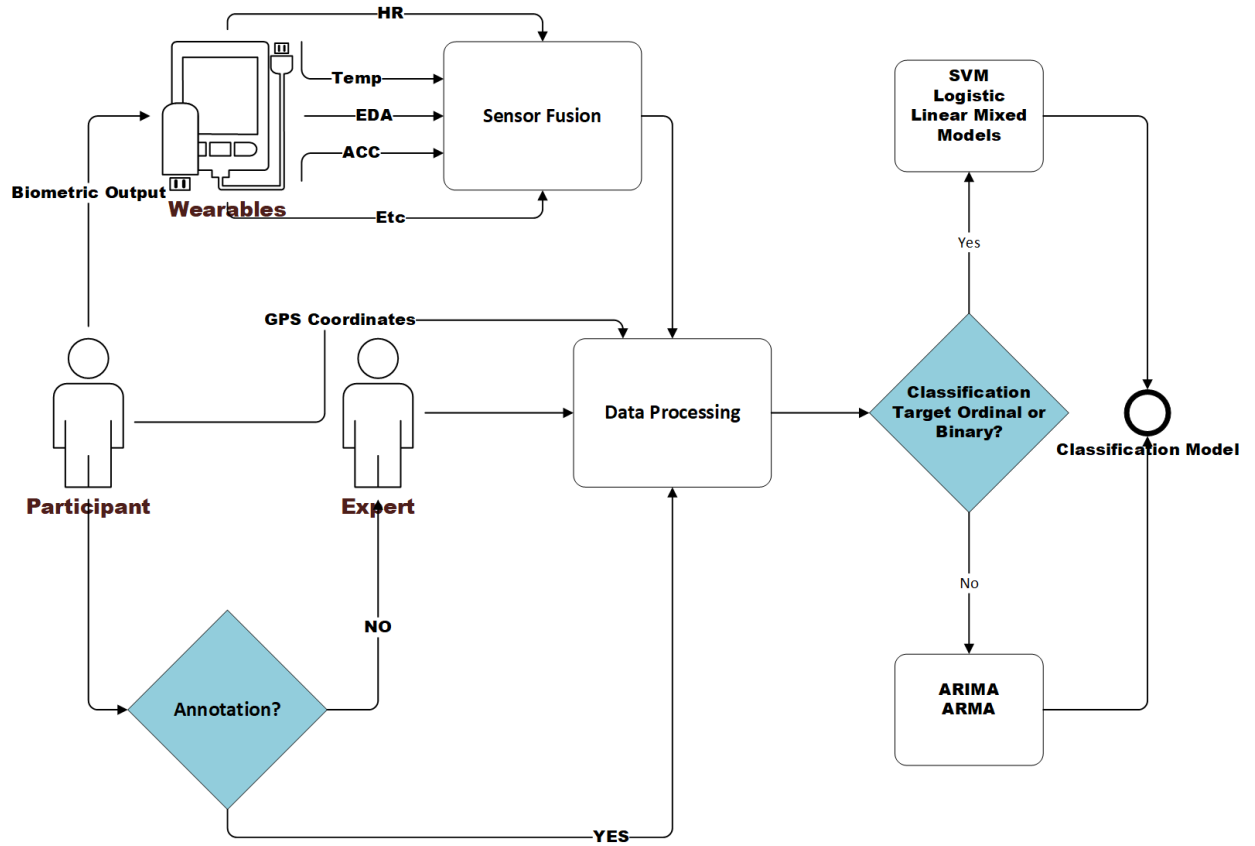
In contrast to ordinal and binary data, nominal data are likely to be collected from users themselves. That is, values that researchers elicit from users directly are holistic, subjective assessments of environmental arousal-induction level per geospatial zone. However, in these experiments, putative arousal annotations by users are treated as additional input variables - channels of input observed over time. A user records these annotations *in situ* during a walk, by using a mobile phone with a slider or hand-held mechanical device to indicate arousal levels to their environment. For example, a device could record a continuous range of values, sampled at a precision of 0 to 1000 numeric values to indicate arousal. Because these variables are sampled using analog devices and converted to digital format, they are treated as continuous-valued¹¹⁷.

4.3.2 Data Processing and Analysis

The following is a proposed design for using wearables and sensors to detect arousal in an affective computing context. First, a participant is provided a wearable that records their biometric data. Second, sensor fusion occurs by combining the input channels heart rate (HR), electrodermal signal (EDA), and temperature (temp) taken from the participant

via the wearable or mobile phone. In addition, there could be other signals, such as an accelerometer or another input of interest. Separately, GPS data are also of interest. Third, data are annotated by the participant either through a Likert survey or another sensor provided to the participant that records a numerical value to indicate stress. Otherwise, a domain expert will examine the data present in the sensor fusion and determine the presence or absence of stress. Fourth, data processing occurs, which prepares data for analysis. The last step considers if the classification target is ordinal/binary/nominal and then applies the appropriate classification method. For example, if data prediction target is binary then it is likely that Support Vector Machines (SVM), logistic regression (logistic), or linear mixed models are appropriate. The final result will be a classification model that can then be used to build an affective intelligent system.

Figure 4.3: *Arousal Detection using Wearables and Sensors in Affective Computing Context*



4.4 Summary

Advances in affective computing and machine learning techniques using wearables and sensors offer a unique opportunity to explore existing and open questions in new ways. This question will be further pursued in chapter 6. For many years, research has focused on the value and benefits of various characteristics of a built environment, including the presence of vegetation, lighting, and public spaces among others. However, many of these studies were conducted in controlled environments and, as a result of developing robust experimental design, were limited in the number of variables and interactions they could test. This dissertation asserts that, by harnessing the capacity of advanced wearables and employing the design methods suggested in the previous section, a carefully designed affective computing approach can help researchers better understand the innumerable variables of design, planning, and their relationship to human health and well-being.

Chapter 5

Binary Classification of Arousal in Built Environments using Machine Learning

5.1 Introduction

5.1.1 Goals

In this chapter, this dissertation proposes using binary classification machine learning techniques, such as logistic regression (LR), random forests (RF), support vector machines (SVM), and multilayer perceptron (MLP), to detect user annotated affect in a built environment. The work in chapter is an extension of the work begun in chapter 3, but with some notable differences. First, this chapter considers user annotation of perceived safety in a built environment as ground truth for the detection of arousal in a built environment. Second, this chapter 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. Like chapter 3, this chapter will address the task of learning a binary

classification model for arousal response in perceived safety.

The motivating goal of this chapter is to develop a methodology and model to classify, predict, and explain the arousal state of participants in a built environment. The viability of demonstrating such a methodology and model indicates the potentiality of developing an intelligent system capable of both classifying and predicting biometric arousal state and automating a process that is traditionally performed by human experts.

5.1.2 Motivations

Research suggests strong correlations and relationships between a physical environment and the influence on the physical and mental well-being of humans^{68;74–77}. There is also evidence that there is a complex relationship between city living, urban upbringing, and the effect of neural stress in humans¹¹⁸. Feelings of safety have been linked as a specific affect regulation system which that has relationships to depression, anxiety, stress, and self-criticism¹¹⁹. Therefore, this chapter and dissertation asserts that affective computing via a machine learning and wearable centric approach is a useful and viable framework in unifying built environments with computer science by providing a methodology to explore these issues. In other words, affective computing can provide a new metric to measure this subjective experience of participants in built environments.

5.1.3 Objectives and Significance

This dissertation proposes to demonstrate that the experimental approach outlined in this chapter indicates the viability of binary classification of participant arousal given feelings of safety in their built environment. This dissertation establishes in chapter 3 that the construction of a machine learning framework for developing an intelligent system in future work is viable. In addition to this, it demonstrates the promise of expert defined annotated data. This chapter wishes to further extend that work by considering user annotation of their experience as ground truth and to determine which machine learning classification approaches are appropriate over others.

This chapter proposes the following novel contributions to the state of the field. First, it is to extend the work in chapter 3 by demonstrating the viability of machine learning algorithms to fit data in a built environment scenario using user defined annotation. Second, it is to explore the performance of logistic regression as compared to the other machine learning algorithms. Third, it is to explore how the findings motivate future work that will be discussed further in Chapter 6.

In conjunction with chapter 3, this dissertation asserts that the detection of arousal using biometric, environmental variables, and user defined annotation is novel and shows promise. This chapter will motivate future work in exploring algorithms and models to both fit, predict, and explain arousal in built environments through binary classification.

5.1.4 Central Hypothesis

This chapter presents the following hypothesis to be tested. That is, to examine the predictive accuracy of the models by fitting machine learning algorithms to the data. Specifically, comparing logistic regression (LR) to RF, SVM, and MLP. Consider the formal hypothesis as follows:

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

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

Here, $A2$ denotes distinct machine learning algorithms, such as RF, SVM, and MLP. The conservative Mann-Whitney test compares the accuracy of both algorithms and makes no parametric distributional assumptions. The test uses $\alpha = 0.05$. In addition, this chapter will examine the AUC (area under the curve) of the algorithms. Model fit and performance is also specifically considered for LR using AIC and McFadden's pseudo R^2 score. Please refer to section 5.5 for results.

5.2 Background and Related Work

5.2.1 Related Work and Wearables

In chapter 3, this dissertation briefly explored the research and related work in estimating arousal using wearable technology. For convenience, this chapter will briefly recap here as well motivation for further research by briefly summarizing the findings in chapter 3. The affective computing approach of detecting arousal using wearable technology is still nascent, but has been around since the 1990s with the dawn of wearable technology. This trend is likely to accelerate given the rise of commercially available technology, such as Empatica, Fitbit, and Garmin. Currently, Empatica is considered medical grade quality and fitbit has been FDA approved^{120;121}. The recent growing sophistication of these biomedical wearable sensors is not only expanding the horizons on what is possible not only in experimental design, data collection, but also in health care and Internet of Things (IoT)^{122;123}.

In this chapter, 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 rest¹²⁴. 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^{125?}. The sensor has also been used in multimodal data collection experiments for mental stress monitoring¹²⁶.

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 algorithms⁶². Binary classification has been explored to correlated physiological and behavioural markers for arousal using wearables¹²⁰, and it is possible to classify panic attacks using wearables and mobile computing⁹³. In chapter 3 of this dissertation, 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^{114;127}.

5.2.2 Random Forests

A classic and very effective machine learning method for binary classification and small data sets is an algorithm called random forests (RF). It is atemporal in nature. 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 or forest is generated, the model uses a vote to combine the tree's predictions¹²⁸. For each tree grown, the following process is followed. First, denote the number in the training set as n such that the algorithm then samples n tuples at random from the original data set. The procedure just described will produce our training sample. Second, suppose there are m input variables such that number p variables out of m are selected at random and the best split on p is used to split the node. It is essential to note that p is held constant during the forest growing. Third, each tree is grown to the largest extent possible and there is no pruning¹²⁹. Finally, training leaves out about approximately 1/3 tuples that are commonly referred to as the out-of-bag (oob) data. This oob is used for classification error as trees are added to the forest. 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 data¹³⁰.

5.2.3 Support Vector Machines

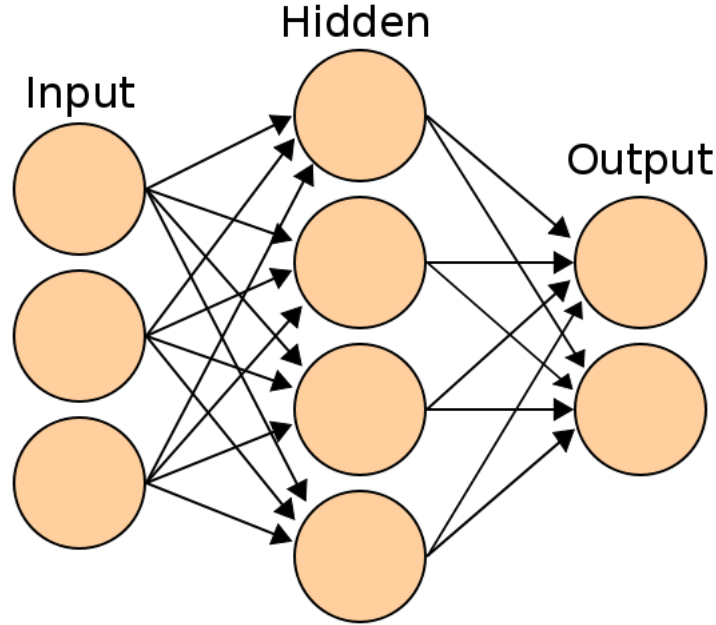
The support vector machine (SVM) is a classic machine learning algorithm that derives its name from the idea that a hyperdimensional plane is compressed down to a plane which divides the hyperdimension into two spaces. In other words, the algorithm partitions the data groups of similar data through a 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 breaches¹³¹.

It is essential to note that SVM is a atemporal binary classifier where the task of the algorithm is to identify a line that separates the data into two classes. The decision boundary is chosen to be the one for which the margin is maximized¹⁰⁰. This can be concisely described as $\min \frac{1}{2} \|\vec{w}\|^2$ such that $y_i(\vec{w} * \vec{x}_i - b) \geq 1$ for all $\vec{x}_i$ where all data points must satisfy the given constraint for the given margins¹²⁹.

5.2.4 Multilayer Perceptrons

One of the classic and most well known artificial neural networks known is the three layer feedforward neural network or multilayer perceptron (MLP), which is the *de facto* standard in ANN topology¹³⁰.

Figure 5.1: *Multilayer Perceptron*



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 layer¹²⁹. Training of the network occurs through adjusting the weights, and the most common way to achieve this is through backpropagation¹³². This

is done by exposing the artificial neural network to data, which then proceed to simulate learning by adjusting weights. 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 function¹³³. In the context of this chapter, artificial neural networks will be used for binary classification.

5.2.5 Logistic Regression

Logistic regression (LR) is a very important, powerful, and spatio-temporal machine learning algorithm that has its origins in statistical learning¹³⁴. The name logistic implies that the relationship is a binary categorical outcome. It is very useful in binary classification and its 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 relationship¹²⁸. Logistic regression is also formally referred 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 follows¹²⁹:

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

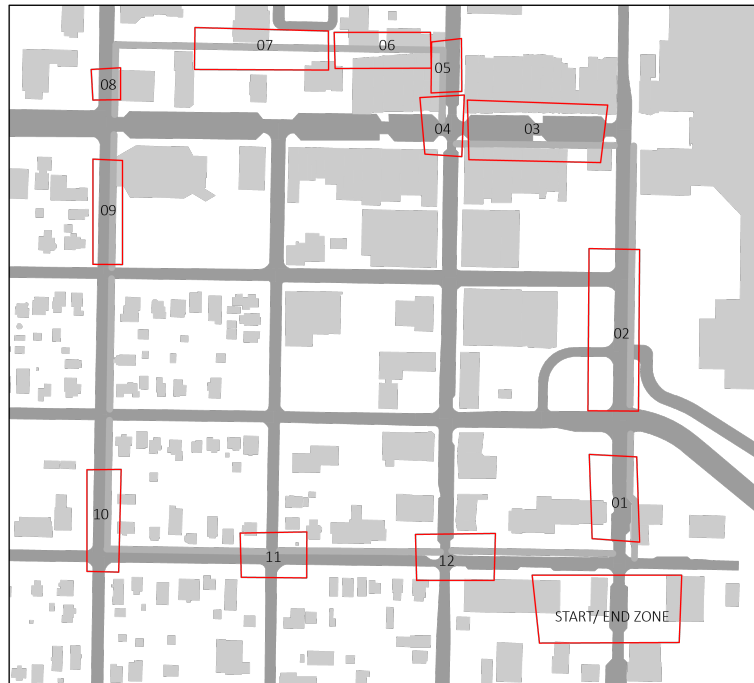
Let $\pi(x)$ denote the probability of the outcome such that $0 \leq \pi(x) \leq 1$ where the logit function $\log\left[\frac{\pi(x)}{1 - \pi(x)}\right]$ has a binomial distribution. The β terms determine the rate of increase or decrease of logistic curve function. That is $\beta > 0$ implies an increase in $\pi(x)$ and $\beta < 0$ implies a decrease in $\pi(x)$. Otherwise the curve is flat¹³⁵.

5.3 Experiment Design: Methodology

5.3.1 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, urban environment, and the likelihood of the environment to invoke an arousal response. Examples include a dark alley, poorly or well lit streets, sidewalks, and calming park like settings.

Figure 5.2: *Urban Built Environment Annotated with Zones*



There were 12 zones chosen above, and each zone is geospatially defined. The zones were named sequentially based on their location in the route as shown above.

In the image above, we can see several urban built environment characteristics and features, such as trees, grass, pavement, buildings, and powerlines. The data set is comprised of 18 participants, which was collected from work done by Parker and Taylor under the direction of Dr. Chamberlain^{69;136}. Each participant was fitted with a Polar V800 and an

Figure 5.3: *Experimental Zone Z7*



Empatica E4. The participant was asked to walk a prearranged route as indicated in the figure below.

Figure 5.4: *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 chapter 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^{69;136}. 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 dissertation 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^{137;138}. Please see section 5.4 and 5.5 for more details.

5.3.2 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 HR and EDA can be an useful and effective methodology in the detection of affective states^{115;139}. 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.

5.3.3 Arousal Prediction Target

This chapter and analysis designate participant annotation to be the ground truth signal of affective state for arousal in the built environment. 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 the 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.

5.3.4 Data Preparation

The data have been cleaned and organized 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 by Taylor Whitaker under the direction of Dr. Brent Chamberlain¹³⁶. The environmental variables include number of street lights, number of trees, number of grass score, walkability score, and what zone the participant was in. In addition, the polar data were tagged with a start and end time to indicate the time the participant began the route and ended the experiment as shown in figure 5.2. 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 5.3.2. The participant arousal response was filtered into a binary classification annotation target as described in section 5.3.3. 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 5.1: *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 5.3.2. 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 domain experts Taylor Whitaker and Dr. Chamberlain¹³⁶.

In addition to the scheme above, additional variables, such as 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.

5.4 Experiment Design: Evaluation Strategy

This section outlines the criteria used in building custom and novel 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 5.2: *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 5.3.4. 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 chapter treated the environmental characteristics as factors for LR. The results are discussed in section 5.5.

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. In addition, results from RF were informative in selecting the environmental features to include in model D.

Second, criteria used for model selection and discrimination was placed upon accuracy and AUC scores. In addition, logistic regression was assessed with Akaike Information Criterion (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 $AIC = 2k - 2\ln(\hat{L})$ such that L is the maximum value of the likelihood function for the model and k is the number of parameters in the model¹⁴⁰. 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 fit¹⁴¹. It can be succinctly described as follows:

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

The model measures the between the likelihood of a full model over a model with just the intercept. The ratio is between 0 and 1. A small ratio between the full model and the intercept model likelihoods subtracted by 1 indicates a better fit than just having a null model. 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 chapter¹⁴².

Third, chapter 3 was primarily concerned with demonstrating the viability of machine learning by fitting the data. This chapter went further by relying 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.

5.4.2 Cross-validation Strategy and Calibration Strategy

This chapter also extends upon the work initially explored in chapter 3 by extending the cross-validation approach. In chapter 3, leave 4 out and leave 3 out cross-validation was implemented. Given the additional 6 participants, this analysis is able to make use of the additional participants and structure of the data to implement leave one out (LOOCV), 2 fold (2FCV), and 3 fold cross-validation (3FCV) strategies. The merit of implementing cross-validation is allowing a more comprehensive picture of model fit and prediction to emerge.

First, train on all participants except leaving one for testing and validation as follows:

Table 5.3: *Leave One Out Cross-validation*

Fold	Train	Test
1	2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18	1
2	1,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18	2
3	1,2,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18	3
4	1,2,3,5,6,7,8,9,10,11,12,13,14,15,16,17,18	4
5	1,2,3,4,6,7,8,9,10,11,12,13,14,15,16,17,18	5
6	1,2,3,4,5,7,8,9,10,11,12,13,14,15,16,17,18	6
7	1,2,3,4,5,6,8,9,10,11,12,13,14,15,16,17,18	7
8	1,2,3,4,5,6,7,9,10,11,12,13,14,15,16,17,18	8
9	1,2,3,4,5,6,7,8,10,11,12,13,14,15,16,17,18	9
10	1,2,3,4,5,6,7,8,9,11,12,13,14,15,16,17,18	10
11	1,2,3,4,5,6,7,8,9,10,12,13,14,15,16,17,18	11
12	1,2,3,4,5,6,7,8,9,10,11,13,14,15,16,17,18	12
13	1,2,3,4,5,6,7,8,9,10,11,12,14,15,16,17,18	13
14	1,2,3,4,5,6,7,8,9,10,11,12,13,15,16,17,18	14
15	1,2,3,4,5,6,7,8,9,10,11,12,13,14,16,17,18	15
16	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,17,18	16
17	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,18	17
18	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17	18

Next, train on all participants except leaving two out as follows:

Table 5.4: *2 Fold Cross-validation*

Fold	Train	Test
1	3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18	1,2
2	1,2,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18	3,4
3	1,2,3,4,7,8,9,10,11,12,13,14,15,16,17,18	5,6
4	1,2,3,4,5,6,9,10,11,12,13,14,15,16,17,18	7,8
5	1,2,3,4,6,7,8,11,12,13,14,15,16,17,18	9,10
6	1,2,3,4,5,7,8,9,10,13,14,15,16,17,18	11,12
7	1,2,3,4,5,6,8,9,10,11,12,15,16,17,18	13,14
8	1,2,3,4,5,6,7,9,10,11,12,13,14,17,18	15,16
9	1,2,3,4,5,6,7,8,10,11,12,13,14,15,16	17,18

Third, this chapter trains on all participants except leaving three participants out for testing:

Table 5.5: *3 Fold Cross-validation*

Fold	Train	Test
1	4,5,6,7,8,9,10,11,12,13,14,15,16,17,18	1,2,3
2	1,2,3,7,8,9,10,11,12,13,14,15,16,17,18	4,5,6
3	1,2,3,4,5,6,10,11,12,13,14,15,16,17,18	7,8,9
4	1,2,3,4,5,6,7,8,9,13,14,15,16,17,18	10,11,12
5	1,2,3,4,6,7,8,9,10,11,12,16,17,18	13,14,15
6	1,2,3,4,5,7,8,9,10,11,12,13,14,15	16,17,18

The models as specified in the beginning of section 5.4 according to criteria outlined in section 5.4.1 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. The details of this analysis can be found in section 5.5.

5.5 Experiment Design: Results

5.5.1 Accuracy and AUC Scores

The accuracy and AUC scores of LR, RF, SVM, and NN will be stated in the tables below. In addition, it is important to note that the analysis also considered two pathological algorithms and models All Negative (AN) and All Positive (AP) where AN estimated no arousal for every tuple and AP estimated arousal for every tuple.

Table 5.6: *LOOCV Accuracy and AUC*

Algorithm	Model	Accuracy	AUC
LR	A	0.5632	0.4977
LR	B	0.7075	0.7239
LR	C	0.6532	0.6643
LR	D	0.7220	0.7357
LR	E	0.7563	0.7842
RF	A	0.5306	0.5151
RF	B	0.7525	0.7797
RF	C	0.7607	0.7792
RF	D	0.7363	0.7582
RF	E	0.6933	0.7236
SVM	A	0.5504	0.4934
SVM	B	0.7373	0.7693
SVM	C	0.5989	0.5081
SVM	D	0.6539	0.6684
SVM	E	0.6223	0.6517
NN	A	0.5459	0.4990
NN	B	0.7208	0.7590
NN	C	0.6101	0.6382
NN	D	0.7241	0.7566
NN	E	0.7053	0.7239

Table 5.7: *2FCV Accuracy and AUC*

Algorithm	Model	Accuracy	AUC
LR	A	0.5399	0.5005
LR	B	0.7548	0.7567
LR	C	0.5179	0.5449
LR	D	0.7585	0.7635
LR	E	0.8096	0.8266
RF	A	0.5201	0.5057
RF	B	0.7944	0.8086
RF	C	0.8041	0.8116
RF	D	0.7790	0.7875
RF	E	0.7263	0.7460
SVM	A	0.5394	0.5020
SVM	B	0.7777	0.7903
SVM	C	0.5835	0.5546
SVM	D	0.6875	0.6909
SVM	E	0.6649	0.6788
NN	A	0.5388	0.5036
NN	B	0.8041	0.8170
NN	C	0.7220	0.7139
NN	D	0.7813	0.7907
NN	E	0.6915	0.7036

Table 5.8: *3FCV Accuracy and AUC*

Algorithm	Model	Accuracy	AUC
LR	A	0.5487	0.4983
LR	B	0.7734	0.7641
LR	C	0.5248	0.5537
LR	D	0.7705	0.7606
LR	E	0.8183	0.8292
RF	A	0.5219	0.4965
RF	B	0.8200	0.8243
RF	C	0.8192	0.8255
RF	D	0.7910	0.7875
RF	E	0.7593	0.7611
SVM	A	0.5493	0.4915
SVM	B	0.7983	0.8046
SVM	C	0.6085	0.5556
SVM	D	0.6900	0.6825
SVM	E	0.6615	0.6587
NN	A	0.5500	0.4941
NN	B	0.8169	0.8226
NN	C	0.6382	0.6534
NN	D	0.7890	0.7822
NN	E	0.7564	0.7579

In general, accuracy improved as the analysis proceeded from leave one out cross-validation to 3 fold cross-validation. This is indicative that the models are likely overfitting the training data for leave one out cross-validation in comparison to the models trained on the data from 3 fold cross-validation that are more robust. It is also interesting to note that RF

and NN, which are atemporal machine learning algorithms performed comparable to the spatiotemporal logistic regression in both accuracy and AUC.

Table 5.9: *AN and AP Accuracy*

Algorithm	Fold	Accuracy
AN	LOOCV	0.5958
AN	2FCV	0.5404
AN	3FCV	0.5643
AP	LOOCV	0.4041
AP	2FCV	0.4595
AP	3FCV	0.4356

The accuracy scores for AN and AP are less than other algorithms and models except for the biometric only defined model A. This difference is evidence that the naive algorithms which either assume total arousal or none across the entire data set is not a good learning strategy. Comparisons of these naive algorithms to LR Model D performance are considered further in section 5.5.2.

5.5.2 Comparisons

In this section, LR model D’s accuracy is compared to the algorithms RF, SVM, NN, AN, and AP 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. The null hypothesis is that the accuracy of LR when compared to algorithms accuracy such as from RF, SVM, NN, AN, and AP is from the same population¹⁴³. This was described more formally in section 5.1.4 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, MLP, AN, or AP. The results of the comparison and tests are given below:

Table 5.10: *LR Model D Accuracy Comparison LOOCV*

Comparison	P-Value
LR and RF	0.5841
LR and SVM	0.0129
LR and NN	0.9129
LR and AN	0.0342
LR and AP	0.0000

Table 5.11: *LR Model D Accuracy Comparison 2FCV*

Comparison	P-Value
LR and RF	0.3401
LR and SVM	0.0027
LR and NN	0.2973
LR and AN	0.0000
LR and AP	0.0000

Table 5.12: *LR Model D Accuracy Comparison 3FCV*

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

First, this dissertation rejects the null hypothesis of accuracy scores between LR and

SVM, LR and AN, and LR and AP 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. Said another way, this dissertation relies on general linear model theory to differentiate LR from the other machine learning algorithms and conduct further inference. The results of this can be read further in the next two sections. That said, the comparisons reveal that LR through model D is a viable machine learning algorithm to detect and predict arousal in participants given biometric and built environment characteristics.

5.5.3 Logistic Model Fit

In this section, this chapter 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 5.13: *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 5.14: *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 5.15: *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

chapter. 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, this chapter will now turn our attention to 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. This chapter will now discuss the explanatory implications of the model in the next section. This chapter will also briefly discuss why model E should be discarded.

5.5.4 Explanatory Model

This chapter and dissertation 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 5.16: *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. In other words, the test conducted is a Wald test in asking if the variable assists in explaining arousal. More formally, it can describe as:

$$H_0 : \beta_{variable} = 0$$

$$H_A : \beta_{variable} \neq 0$$

Where *variable* refers to the specific coefficient being tested. For example, Norm HR is highly significant in contributing to explaining arousal in the model at a p-value of nearly 0. Before this chapter proceeds further, let this dissertation 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.

Special attention should be paid to whether or not the coefficient of a variable is greater than 0 or less than 0. The LR algorithm will report the probability between 0 and 1 that an arousal event given the input given it. If a coefficient is less than 0, then it contributes to the likelihood that the participant will have less chance of arousal. Conversely, if the coefficient is greater than 0, then it contributes to the likelihood of the model reporting a higher chance of arousal. Said differently, arousal here is associated with a measure of safety. The model suggests that higher heart rate and presences 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.

There are some important caveats to mention. This chapter and dissertation asserts 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 chapter and dissertation do establish a connection between a participants biometrics, environmental variables, and the detection of their arousal affects via machine learning. Please see chapter 6 for further discussion of future work and implications of the explanatory model.

5.6 Summary

Many machine learning researchers have noted that there is no single best model that works optimally for all kinds of problems¹²⁸. However, the results in this chapter certainly suggest that LR among a few other 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.

Chapter 6

Summary and Future Work

6.1 Summary

The results in chapter 3 demonstrate the feasibility of learning arousal state from a combination of biometric data and built environment data using expert defined annotated data. This chapter also demonstrates the viability for prediction by promising AUC scores. The sample size was small, and statistical inference could not be made to differentiate which method might be better. The promising results in chapter 3 in conjunction with the continuing evolution in affective computing and machine learning techniques using wearables and sensors offer unique opportunities in exploring existing and open questions in new ways. Traditional studies before affective computing have relied on non-machine learning summary statistics and laboratory setting studies, including built environments. By proposing a methodology to fuse mobile wearables and user annotation using a machine learning centric approach to detect affect, this dissertation asserts that a carefully designed experiment can better help researchers understand the innumerable variables and their influence on affect in a built environment. The results in chapter 5 suggest that logistic regression provides a useful approach and methodology in the detection of affect in a built environment. In addition, there are a few other machine learning algorithms that also showed promise, such as Random Forests and Multi-Layer Perceptrons. The accuracy, AUC, AIC, Chi-square tests, and pseudo R^2

provided a useful toolkit in assessing model performance. First, the Chi-square tests revealed that the tests were statistically useful. Second, the AIC score provided a useful measure in suggesting model D as an appropriate model over others. The pR^2 , Wald tests, accuracy and AUC indicated that LR model D fit the data well, had predictive abilities, and contained information in the coefficients that contributed to further understanding of the relationship between biometric data, environmental characteristics, and arousal in a built environment. Wolpert famously said that there is no single best model that works optimally for all kinds of problems¹⁴⁴, but this dissertation asserts there are several promising models for the detection of affect in built environments.

6.2 Conclusions

This dissertation asserts three novel contributions to the state of the field. First, this dissertation proposed a design methodology for affective computing in built environments. In Chapter 3, it was demonstrated that machine learning is a viable approach in the detection of arousal defined by expert annotation. In Chapter 4, a design methodology for the detection of affect in a built environment that relies on user annotation and sensor fusion was proposed. Results in chapter 5 showed promising results and several potential avenues for future work and further refinement of design methodology for the detection of affect in a built environment. Please see the discussion of future work for further details.

Second, this dissertation asserts that normalized HR, EDA, and environmental characteristics, such as walkability, grass, trees, electrical transformers, poles, number of power lines, and lights are pertinent variables from built environments and sensors to detect user arousal. The sample size of 18 participants necessitates the need for larger scale experiments to further investigate pertinent biographical and environmental variables useful in the detection of arousal from participants in built environments.

Third, this dissertation asserts that a machine learning centric approach to built environments enables affective intelligence in built environments. The work explored in this dissertation has established the viability of machine learning to detect both expert and user

defined arousal. The binary classification of arousal by logistic regression suggests a baseline for current and future research in this domain. In addition, the auto-regressive features that GLMM used in machine learning is suggestive of future work with a focus of auto-regressive and spatio-temporal machine learning algorithms with continuous annotation from a user as a prediction target. The surprising success of atemporal methods such as Random Forests and Multilayer Perceptron suggests future potential in new kinds of affective computing experiments in built environments.

In conclusion, affective computing is the proposed framework and umbrella to combine future research in built environment and affective response via a machine learning centric approach.

6.3 Limitations

The research in this dissertation has generated several open questions during this investigation which could not be fully pursued due to the the following challenges. The sample size of the data was limited to 18 participants. It is desirable that future work exceeds this number and seek as many participants as possible. In addition, the sample size should also reflect a more diverse population of participants both in age, gender, race, and culture. In parallel to this, a plethora and diversity of different urban and rural built environments should be explored. In addition, more sophisticated forms of sensor fusion should be pursued such real-time feature extraction of environmental characteristics using machine learning vision. The collection of audio information was attempted, but not all participants were consistent in the verbal annotation, indicating a new method of real time annotation should be pursued.

6.3.1 Open Questions

There are several major open questions that remain. This dissertation will organize them as follows. The first paragraph will discuss open questions associated with design methodology, built environment characteristics, and some of the wider issues that are also associated with

affective computing. The second paragraph will address questions associated with modeling and machine learning.

There are many open questions that still remain worth pursuing that involve design, methodology, built environments, and affective computing as follows: 1) Several major questions remain: 1) Can we differentiate effects between various design characteristics and identify strong correlation between those and arousal? 2) Does an affective computing approach offer additional insight that traditional research methods (and related findings) for studying built environments have not addressed? 3) Can affective computing be used to understand the difference between effects from environmental characteristics, social interactions and if those are mutually exclusive in different contexts? 4) Are there particular cultural differences or individual circumstances that influence results? 5) What patterns exist between daily behaviors and built environments, and are they influencing one another? 6) What is the most ethical way to commercialize affective systems designed for built environment spaces?

There are many open questions surrounding modeling and machine learning in built environments: 1) Is modeling for built environments via machine learning most appropriate by using temporal algorithms, atemporal algorithms, or both? 2) What are the additional biographical information that may serve useful in building a more nuanced and generalize model? 3) How can models be extended beyond the detection of arousal and detect contextual affect, such as happiness, sadness, anger, and etc in a built environment? 4) What is the ideal window for an experimental zone for binary classification of affect in a built environment?

Arguably, the most important step moving forward is to expand collaborations and the interdisciplinary approach to studying built environments in an affective computing context. Researchers and industry should seek to address these questions by working together. Based on this dissertation, it is hoped that future research will address these questions in the long term, but the first aim is to develop reproducible results by aiding interested researchers to use the same design methodologies and approach as developed in this dissertation while also relying on the ever increasing power of wearables and capabilities of affective computing.

6.4 Future Work

There are several areas for future work. First, the development of a sensor to collect and process real time user annotation in a built environment. This sensor will have to be validated and tested in several experiments. Once this sensor is completed and validated, it will have to be deployed in actual built environments. Second, this dissertation focused on discrete experimental zones. Future work will focus on the processing of continuous zones that can be aggregated based on environmental characteristics, such as trees, electrical poles and lines, buildings, and other relevant environmental variables. Third, future studies should focus on the role that time and weather might have an influencing affect in a built environment. In longer term work, continuing work should focus on evolving this multi-sensor approach to arousal state detection and prediction in a built environment. Future experiments should be pursued with as many participants as possible.

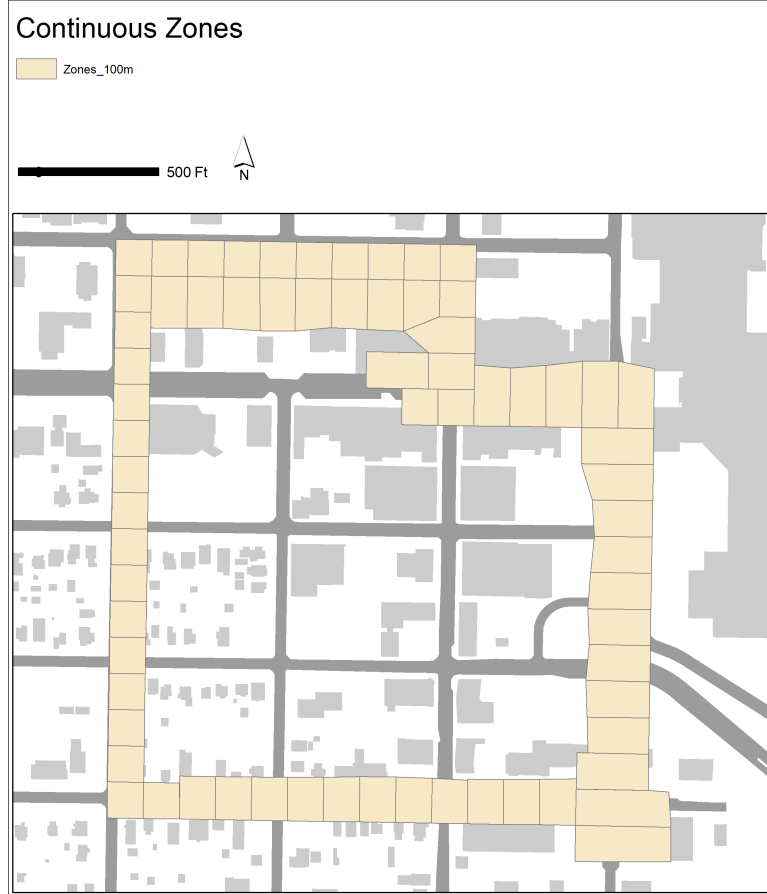
6.4.1 Short-Term

Chamberlain and Whitaker recently conducted a new experiment with 10 more participants. Using the results and approach outlined in this dissertation, it will be of interest to examine the new participants to investigate if time of day influences the detection of arousal in a built environment. In addition, information about the weather, such as temperature and weather, may also be used.

This dissertation focused on binary classification of arousal based on expert defined arousal and user defined arousal. The experimental zones used in this experiment were selected specially by domain experts and discrete. Future work will focus on the continuous zones, where each zone will represent an aggregate of environmental characteristics, such as grass, shrubs, trees, poles, lines, buildings, and walkability score.

In addition, the development of sensors with the potential for real time annotation will also be explored. It will be necessary to first validate these sensor's performance in the detection of arousal in built environments. After the verification is completed, the sensors will be deployed in real built environments.

Figure 6.1: *Continuous Zones*



It is also of interest to pursue the multi-modal pursuit of affect by using empathica, visual, and audio data collection. Specifically, the pursuit of arousal detection from users watching high quality video scenes of various built environments. In addition, relying on user annotation to report affect as the prediction target. This research is a natural extension of exploring the potential atemporal nature of the interaction between built environments.

6.4.2 Long-Term

In the long term, continuous zones, real time annotation, and sensor fusion should facilitate the collection of data on larger and larger scales. Models should be generalized to detect affect

in various built environments and contextual situations, such as day, night, cold, or warm weather. The models should also be robust to detect core affect across cross-cultural built environments. In addition, the models must evolve to address diversity inherent in the human experience, such as age, gender, race, health, and urban built environment preferences. In summary, experiments also must be designed to increase sample size, balance gender, and increase diversity as much as possible.

Research is needed to further study the influence of built environment characteristics on biometric signals like HR and EDA. The state of the field is to rely either on user or expert annotation to report arousal in data. Future research must focus on using machine learning to detect patterns and signals in biometrics that can indicate arousal without the need for user or expert annotation. In other words, it is of interest to develop algorithms that can detect potential arousal of participants based on biometric output given built environment characteristics without supervised labeling of the data.

In addition, there are several possibilities for the development of novel algorithmic approaches to the detection of affect in built environments. As experiments grow increasingly more sophisticated, it is likely that logistic regression will continue to evolve in sophistication and potentially necessitate the evolution into general linear mixed models approach (GLMM). There is also potential of the hybridization of artificial neural networks in conjunction with general linear mixed models, both increasing the accuracy of the estimation techniques used by GLMM and accounting for the random and fixed effects in built environment experiments. Real time annotation will necessitate the need to pursue time series and deep learning algorithms. The same is likely true for biometric signal data, such as IBI. In both scenarios outlined above, it is likely that the algorithms could be evolved to work with methodologies pioneered by Dr. Hsu on Self-Organized-Expert Modular for Classification of Spatiotemporal sequences (SMNET) hybridized to work with deep learning and GLMM¹⁴⁵.

The contextual recognition of a plethora of core affective states in humans given their built environments is likely a lifetime pursuit. Current, near-term, and even some long-term research endeavors will focus on the detection of arousal in built environments while enhancing the experimental design, methodologies, sensors, algorithms, and capabilities and

abilities of models to detect arousal in built environments. However, the contextual recognition of emotions, such as surprise, interest, joy, rage, fear, disgust, shame, and anguish that research has shown to be core affect and cross-cultural¹⁴⁶ is the ultimate viable and long term strategy. The capability of this enhanced affective intelligence in built environments provides the opportunity to close the circle with human designed built environments by providing our built environments the ability to understand us for the betterment of human health and well-being.

Chapter 7

Code References

A significant aspect of this dissertation research was the development and construction of custom code. The code is shared on a public Github repository so that interested readers can understand the source code. Requests for the raw or processed data can be referred to Dr. Brent Chamberlain.

The following is a brief description and explanation of the source code on Github. The repository is called `AffectiveIntelligenceBuiltEnvironments` and can be accessed at the following url: <https://github.com/heathyates/AffectiveIntelligenceBuiltEnvironments>. The code is modular and function based which should enable easy maintenance and facilitate with further development. The github repository is organized into 4 folders that contains R scripts that accomplish the following:

- **0. Preprocess Data** - Takes the raw data and organizes it for cleaning and processing.
- **1. Clean and Process Data** - Cleans the raw E4 data and process Polar + Environments data
- **2. Post-Cleaning Processing** - Removes data before and after start/end tag and removes a few outliers
- **3. Generate Cross-Validation Data** - Does LOOCV, 2FCV, and 3FCV

- **4. Analysis - Spatiotemporal** - Named because the cross-validation was processed by participant ID and Time and fed into algorithms such as Logistic Regression, Random Forests, SVM, and Multi-Layer Perceptron.

The ordering above indicates the expected order in which to run the scripts. For example, it is necessary to run the preprocessing script before proceeding with cleaning and processing the data. It is possible that additional future work will focus on porting this source code into Python to work better with arcGIS. However, that is beyond the scope of this dissertation and current research. This code has also been designed so that it can be further developed to work with automation scripts such as Power Shell or Bash.

Bibliography

- [1] Rosalind W Picard and Roalind Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997.
- [2] Jianhua Tao and Tieniu Tan. Affective computing: A review. In *International Conference on Affective computing and intelligent interaction*, pages 981–995. Springer, 2005.
- [3] RM Page. Man-machine coupling-2012 ad. *Proceedings of the IRE*, 50(5):613–614, 1962.
- [4] Rosalind W Picard. Building hal: Computers that sense, recognize, and respond to human emotion. In *Photonics West 2001-Electronic Imaging*, pages 518–523. International Society for Optics and Photonics, 2001.
- [5] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, 23(10):1175–1191, 2001.
- [6] David H Gustafson, Tara E Palesh, Rosalind W Picard, Paul E Plsek, Lynne Maher, and Victor A Capoccia. Automating addiction treatment: Enhancing the human experience and creating a fix for the future. *Studies in health technology and informatics*, 118:186–206, 2005.
- [7] Rana El Kaliouby, Alea Teeters, and Rosalind W Picard. An exploratory social-emotional prosthetic for autism spectrum disorders. In *Wearable and Implantable Body Sensor Networks, 2006. BSN 2006. International Workshop on*, pages 2–pp. IEEE, 2006.

- [8] Rosalind W Picard. Toward machines with emotional intelligence. In *ICINCO (Invited Speakers)*, pages 29–30, 2004.
- [9] Rosalind W Picard and Jennifer Healey. Affective wearables. *Personal Technologies*, 1(4):231–240, 1997.
- [10] Alex P Pentland. *Wearable intelligence*. Scientific American, Incorporated, 1998.
- [11] Jennifer Anne Healey. *Wearable and automotive systems for affect recognition from physiology*. PhD thesis, Massachusetts Institute of Technology, 2000.
- [12] Rosalind W Picard and W Rosalind. Toward agents that recognize emotion. *VIVEK-BOMBAY-*, 13(1):3–13, 2000.
- [13] Rosalind W Picard, Seymour Papert, Walter Bender, Bruce Blumberg, Cynthia Breazeal, David Cavallo, Tod Machover, Mitchel Resnick, Deb Roy, and Carol Strohecker. Affective learning—a manifesto. *BT technology journal*, 22(4):253–269, 2004.
- [14] John T Cacioppo and Louis G Tassinary. Inferring psychological significance from physiological signals. *American psychologist*, 45(1):16, 1990.
- [15] John T Cacioppo, Louis G Tassinary, and Alan J Fridlund. The skeletomotor system. 1990.
- [16] Michael E Dawson, Anne M Schell, and Diane L Filion. The electrodermal system. *Handbook of psychophysiology*, 2:200–223, 2007.
- [17] John T Cacioppo and Louis G Tassinary. *Principles of psychophysiology: Physical, social, and inferential elements*. Cambridge University Press, 1990.
- [18] Rosalind W Picard. Emotion research by the people, for the people. *Emotion Review*, 2010.
- [19] J Rick Turner and Marc Gellman. *Encyclopedia of Behavioral Medicine*. Springer, 2013.

- [20] Roof and Oleru. Public health: Seattle and king county’s push for the built environment. *J Environ Health*, 71:24–27, 2008.
- [21] Margaret S Clark. A role for arousal in the link between feeling states, judgments, and behavior. In *Affect and Cognition: 17th Annual Carnegie Mellon Symposium on Cognition*, page 263. Psychology Press, 2014.
- [22] Rosalind W Picard. Human-computer coupling. *Proceedings of the IEEE*, 86(8):1803–1807, 1998.
- [23] Rosalind W Picard. Affective computing for hci. In *HCI (1)*, pages 829–833, 1999.
- [24] Carson Reynolds and Rosalind W Picard. Designing for affective interactions. In *Proceedings from the 9th International Conference on Human-Computer Interaction*, page 6, 2001.
- [25] Daniel Goleman. *Emotional intelligence*. Bantam, 2006.
- [26] Jocelyn Riseberg, Jonathan Klein, Raul Fernandez, and Rosalind W Picard. Frustrating the user on purpose: Using biosignals in a pilot study to detect the user’s emotional state. In *CHI 98 Cconference Summary on Human Factors in Computing Systems*, pages 227–228. ACM, 1998.
- [27] K Liu and Rosalind W Picard. Embedded empathy in continuous, interactive health assessment. In *CHI Workshop on HCI Challenges in Health Assessment*, volume 1, page 3. Citeseer, 2005.
- [28] Timothy Bickmore, Amanda Gruber, and Rosalind Picard. Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient education and counseling*, 59(1):21–30, 2005.
- [29] Rosalind W Picard and Shaundra Bryant Daily. Evaluating affective interactions: Alternatives to asking what users feel. In *CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches*, pages 2119–2122. ACM New York, NY, 2005.

- [30] Jennifer Healey and Rosalind Picard. Digital processing of affective signals. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 6, pages 3749–3752. IEEE, 1998.
- [31] Jennifer Healey and Rosalind W Picard. Startlecam: A cybernetic wearable camera. In *Wearable Computers, 1998. Digest of Papers. Second International Symposium on*, pages 42–49. IEEE, 1998.
- [32] Jocelyn Scheirer, Raul Fernandez, and Rosalind W Picard. Expression glasses: a wearable device for facial expression recognition. In *CHI'99 Extended Abstracts on Human Factors in Computing Systems*, pages 262–263. ACM, 1999.
- [33] RW Picard and E Vyzas. ^aoffline and online recognition of emotion expression from physiological data, ^o emotion-based agent architectures workshop notes, int'l conf. In *Autonomous Agents*, pages 135–142, 1999.
- [34] Jennifer Healey, Justin Seger, and Rosalind Picard. Quantifying driver stress: Developing a system for collecting and processing bio-metric signals in natural situations. *Biomedical sciences instrumentation*, 35:193–198, 1999.
- [35] Rosalind W Picard. Affective medicine: Technology with emotional intelligence. *Studies in health technology and informatics*, pages 69–84, 2002.
- [36] Jennifer Healey and Rosalind Picard. Smartcar: detecting driver stress. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 4, pages 218–221. IEEE, 2000.
- [37] Ashish Kapoor and Rosalind W Picard. A real-time head nod and shake detector. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–5. ACM, 2001.
- [38] Katharine Blocher and Rosalind W Picard. Affective social quest. In *Socially intelligent agents*, pages 133–140. Springer, 2002.

- [39] Rosalind W Picard and C Du. Monitoring stress and heart health with a phone and wearable computer. *Motorola Offspring Journal*, 1:14–22, 2002.
- [40] Ashish Kapoor and Rosalind W Picard. Real-time, fully automatic upper facial feature tracking. In *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pages 10–15. IEEE, 2002.
- [41] Raul Fernandez and Rosalind W Picard. Modeling drivers’ speech under stress. *Speech communication*, 40(1):145–159, 2003.
- [42] Carson Reynolds and Rosalind Picard. Affective sensors, privacy, and ethical contracts. In *CHI’04 Extended Abstracts on Human Factors in Computing Systems*, pages 1103–1106. ACM, 2004.
- [43] Christopher R Wren and Carson J Reynolds. Minimalism in ubiquitous interface design. *Personal and Ubiquitous Computing*, 8(5):370–373, 2004.
- [44] Timothy W Bickmore and Rosalind W Picard. Towards caring machines. In *CHI’04 extended abstracts on Human factors in computing systems*, pages 1489–1492. ACM, 2004.
- [45] Carson Jonathan Reynolds. *Adversarial uses of affective computing and ethical implications*. PhD thesis, Massachusetts Institute of Technology, 2005.
- [46] Kevin Larson and R Picard. The aesthetics of reading. In *Appears in Human-Computer Interaction Consortium Conference, Snow Mountain Ranch, Fraser, Colorado*, 2005.
- [47] 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.
- [48] Marc Strauss, Carson Reynolds, Stephen Hughes, Kyoung Park, Gary McDarby, and Rosalind Picard. The handwave bluetooth skin conductance sensor. *Affective computing and intelligent interaction*, pages 699–706, 2005.

- [49] Ashish Kapoor and Rosalind W Picard. Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 677–682. ACM, 2005.
- [50] Ashish Kapoor, Winslow Burleson, and Rosalind W Picard. Automatic prediction of frustration. *International journal of human-computer studies*, 65(8):724–736, 2007.
- [51] Mohammed E Hoque. Analysis of speech properties of neurotypicals and individuals diagnosed with autism and down. In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*, pages 311–312. ACM, 2008.
- [52] Ming-Zher Poh, Nicholas C Swenson, and Rosalind W Picard. Comfortable sensor wristband for ambulatory assessment of electrodermal activity. In *1st Biennial Conference of the Society for Ambulatory Assessment, Greifswald, Germany*, 2009.
- [53] E Hedman, M Eckhardt, Ming-Zher Poh, M. S. Goodwin, L. J. Miller, B. Brett-Green, S. A. Schoen, D. M. Nielsen, and Rosalind W Picard. Heart rate variability and electrodermal activity in children with atypical sensory processing: Exploratory pattern analysis. In *International Journal of Human-Computer Studies*, 2009.
- [54] Ming-Zher Poh, Nicholas C Swenson, and Rosalind W Picard. A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE transactions on Biomedical engineering*, 57(5):1243–1252, 2010.
- [55] Richard Ribon Fletcher, Kelly Dobson, Matthew S Goodwin, Hoda Eydgahi, Oliver Wilder-Smith, David Fernholz, Yuta Kuboyama, Elliott Bruce Hedman, Ming-Zher Poh, and Rosalind W Picard. icalm: Wearable sensor and network architecture for wirelessly communicating and logging autonomic activity. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):215–223, 2010.
- [56] Richard R Fletcher, Ming-Zher Poh, and Hoda Eydgahi. Wearable sensors: opportunities and challenges for low-cost health care. In *Engineering in Medicine and Biology*

- Society (EMBC), 2010 Annual International Conference of the IEEE*, pages 1763–1766. IEEE, 2010.
- [57] Ming-Zher Poh, Daniel McDuff, and Rosalind Picard. A medical mirror for non-contact health monitoring. In *ACM SIGGRAPH 2011 Emerging Technologies*, page 2. ACM, 2011.
 - [58] Mohammed Hoque, Louis-Philippe Morency, and Rosalind Picard. Are you friendly or just polite?—analysis of smiles in spontaneous face-to-face interactions. *Affective Computing and Intelligent Interaction*, pages 135–144, 2011.
 - [59] Ming-Zher Poh, Kyunghye Kim, Andrew Goessling, Nicholas Swenson, and Rosalind Picard. Cardiovascular monitoring using earphones and a mobile device. *IEEE Pervasive Computing*, 11(4):18–26, 2012.
 - [60] Javier Hernandez, Daniel McDuff, Rich Fletcher, and Rosalind W Picard. Inside-out: Reflecting on your inner state. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on*, pages 324–327. IEEE, 2013.
 - [61] Daniel McDuff, Rana El Kaliouby, Evan Kodra, and Rosalind Picard. Measuring voter’s candidate preference based on affective responses to election debates. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 369–374. IEEE, 2013.
 - [62] Akane Sano and Rosalind W Picard. Stress recognition using wearable sensors and mobile phones. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 671–676. IEEE, 2013.
 - [63] Maurizio Garbarino, Matteo Lai, Dan Bender, Rosalind W Picard, and Simone Tognetti. Empatica e3—a wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *Wireless Mobile Communication and Health-*

- care (Mobihealth)*, 2014 *EAI 4th International Conference on*, pages 39–42. IEEE, 2014.
- [64] Daniel McDuff, Sarah Gontarek, and Rosalind Picard. Remote measurement of cognitive stress via heart rate variability. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 2957–2960. IEEE, 2014.
- [65] Javier Hernandez Rivera, Daniel Leithinger, David Adley Mellis, and Roy Shilkrot. Media arts and sciences-ph. d./sc. d. research and teaching output of the mit community.
- [66] Ming-Zher Poh, Tobias Loddenkemper, Claus Reinsberger, Nicholas C Swenson, Shubhi Goyal, Mangwe C Sabtala, Joseph R Madsen, and Rosalind W Picard. Convulsive seizure detection using a wrist-worn electrodermal activity and accelerometry biosensor. *Epilepsia*, 53(5), 2012.
- [67] Natasha Jaques, Sara Taylor, Akane Sano, Rosalind Picard, et al. Predicting tomorrow’s mood, health, and stress level using personalized multitask learning and domain adaptation. In *IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*, pages 17–33, 2017.
- [68] G. W. Evans. Environmental stress. *CUP Archive*, 1984.
- [69] R Parker. Your environment and you: Investigating stress triggers and characteristics of the built environment. 2016.
- [70] Jeremy P Jamieson, Matthew K Nock, and Wendy Berry Mendes. Mind over matter: Reappraising arousal improves cardiovascular and cognitive responses to stress. *Journal of Experimental Psychology: General*, 141(3):417, 2012.
- [71] Sigrid Adriaenssens, Philippe Block, Diederik Veenendaal, and Chris Williams. *Shell structures for architecture: form finding and optimization*. Routledge, 2014.

- [72] Mariana Popescu, Matthias Rippmann, Tom Van Mele, and Philippe Block. Automated generation of knit patterns for non-developable surfaces. In *Humanizing Digital Reality*, pages 271–284. Springer, 2018.
- [73] Thies Lindenthal. Estimating supply elasticities for residential real estate in the united kingdom. 2018.
- [74] 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.
- [75] G. W. Evans. The built environment and mental health. *Journal of Urban Health*, 30: 536–555, 2003.
- [76] 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.
- [77] 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.
- [78] R. S. Ulrich. Natural versus urban scenes. *Environment and Behavior*, 13:523, 1981.
- [79] 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.
- [80] 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.
- [81] 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.
- [82] 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.
- [83] R. J. Jackson. The impact of the built environment on health: an emerging field. *American Public Health Association*, 2003.
- [84] R. S. Ulrich. Aesthetic and affective response to natural environment. *Behavior and the natural environment*, pages 58–125, 1983.
- [85] R. Kaplan and S. Kaplan. The experience of nature: A psychological perspective. *CUP Archive*, 1989.
- [86] S. Kaplan. The restorative benefits of nature: Toward an integrative framework. *Journal of Environmental Psychology*, 15:169–182, 1995.
- [87] Roger S Ulrich, Robert F Simons, Barbara D Losito, Evelyn Fiorito, Mark A Miles, and Michael Zelson. Stress recovery during exposure to natural and urban environments. *Journal of environmental psychology*, 11(3):201–230, 1991.
- [88] Y. Tsunetsugu, J. Lee, B. J. Park, L. Tyrväinen, 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.
- [89] 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.
- [90] J Hernandez Rivera. *Towards wearable stress measurement (Doctoral Dissertation)*. MIT Press, Cambridge, Massachusetts, 2015.

- [91] 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.
- [92] 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.
- [93] 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.
- [94] R. Hoyt and P. Karl. The future of wearable tech. *US Army*. Retrieved from <https://www.army.mil/article/161761/>, February 2016.
- [95] 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.
- [96] 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.
- [97] J. F. Thayer and R. D. Lane. Claude bernard and the heart–brain connection: Further elaboration of a model of neurovisceral integration. *Neuroscience Biobehavioral Reviews*, 33(2):81–88, 2009.
- [98] B. S. McEwen. Physiology and neurobiology of stress and adaptation: Central role of the brain integration. *Physiology Review*, 87:873–904, 2007.
- [99] C Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, New York, 2006.

- [100] K. P. Murphy. *Machine Learning*. MIT Press, Cambridge, Massachusetts, 2012.
- [101] D. R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society*, pages 215–242, 1958.
- [102] N. E. Brewslo and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, 1993.
- [103] Christopher Alexander. *A pattern language: towns, buildings, construction*. Oxford university press, 1977.
- [104] Francis DK Ching. *Architecture: Form, space, and order*. John Wiley & Sons, 2014.
- [105] Don Alexander. Physical determinism, modernism and mental health. *Environments*, 35(3):81, 2008.
- [106] Akane Sano and Rosalind W Picard. Stress recognition using wearable sensors and mobile phones. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 671–676. IEEE, 2013.
- [107] Simon Ollander, Christelle Godin, Aurélie Campagne, and Sylvie Charbonnier. A comparison of wearable and stationary sensors for stress detection. In *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on*, pages 004362–004366. IEEE, 2016.
- [108] Fatema El-Amrawy and Mohamed Ismail Nounou. Are currently available wearable devices for activity tracking and heart rate monitoring accurate, precise, and medically beneficial? *Healthcare informatics research*, 21(4):315–320, 2015.
- [109] Stanley Schachter and Jerome Singer. Cognitive, social, and physiological determinants of emotional state. *Psychological review*, 69(5):379, 1962.
- [110] Ming-Zher Poh, Nicholas C Swenson, and Rosalind W Picard. Comfortable sensor wristband for ambulatory assessment of electrodermal activity. In *1st Biennial Conference of the Society for Ambulatory Assessment, Greifswald, Germany*, 2009.

- [111] Scott Weich, Martin Blanchard, Martin Prince, Elizabeth Burton, BOB Erens, and Kerry Sproston. Mental health and the built environment: Cross-sectional survey of individual and contextual risk factors for depression. *The British Journal of Psychiatry*, 180(5):428–433, 2002.
- [112] Abraham Wandersman and Maury Nation. Urban neighborhoods and mental health: Psychological contributions to understanding toxicity, resilience, and interventions. *American Psychologist*, 53(6):647, 1998.
- [113] AJ Milam, CDM Furr-Holden, and PJ Leaf. Perceived school and neighborhood safety, neighborhood violence and academic achievement in urban school children. *The Urban Review*, 42(5):458–467, 2010.
- [114] 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, 2017.
- [115] 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.
- [116] Erling B Andersen. Discrete statistical models with social science applications. Technical report, 1980.
- [117] Gaël Laurans, Pieter MA Desmet, and Paul Hekkert. The emotion slider: A self-report device for the continuous measurement of emotion. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–6. Ieee, 2009.
- [118] 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.

- [119] 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.
- [120] Maurizio Garbarino, Matteo Lai, Dan Bender, Rosalind W Picard, and Simone Tognetti. Empatica e3—a 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.
- [121] Casey Erdmier, Jason Hatcher, and Michael Lee. Wearable device implications in the healthcare industry. *Journal of medical engineering & technology*, 40(4):141–148, 2016.
- [122] 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.
- [123] 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.
- [124] 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.
- [125] 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.
- [126] 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.
- [127] 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, 2017.
 - [128] Christian Robert. Machine learning, a probabilistic perspective, 2014.
 - [129] M Bishop Christopher. *PATTERN RECOGNITION AND MACHINE LEARNING*. Springer-Verlag New York, 2016.
 - [130] Brett Lantz. *Machine learning with R*. Packt Publishing Ltd, 2015.
 - [131] Lipo Wang. *Support vector machines: theory and applications*, volume 177. Springer Science & Business Media, 2005.
 - [132] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533, 1986.
 - [133] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.
 - [134] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.
 - [135] Alan Agresti. *Logistic regression*. Wiley Online Library, 2002.
 - [136] T Whitaker. Linking affect and the built environment using mobile sensors and geospatial analysis. 2018.
 - [137] David Meyer. Support vector machines: The interface to libsvm in package e1071. 2004.

- [138] Brian Ripley, William Venables, and Maintainer Brian Ripley. Package ‘nnet’. *R package version*, pages 7–3, 2016.
- [139] 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.
- [140] John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.
- [141] Daniel McFadden. The measurement of urban travel demand. *Journal of public economics*, 3(4):303–328, 1974.
- [142] David A Hensher and Peter R Stopher. *Behavioural travel modelling*. Taylor & Francis, 1979.
- [143] James J Higgins. Introduction to modern nonparametric statistics. 2003.
- [144] David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.
- [145] William H Hsu and Sylvian R Ray. A new mixture model for concept learning from time series. In *In Proceedings of the 1998 Joint AAAI-ICML Workshop on AI Approaches to Time Series Problems (Technical Report WS-98-07)*. Citeseer, 1998.
- [146] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.