University of Memphis
University of Memphis Digital Commons

Electronic Theses and Dissertations

5-26-2023

Continuous Estimation of Smoking Lapse Risk from Noisy Wrist Sensor Data Using Sparse and Positive-Only Labels

Md Azim Ullah

Follow this and additional works at: https://digitalcommons.memphis.edu/etd

Recommended Citation
Ullah, Md Azim, "Continuous Estimation of Smoking Lapse Risk from Noisy Wrist Sensor Data Using Sparse and Positive-Only Labels" (2023). *Electronic Theses and Dissertations*. 3083.
https://digitalcommons.memphis.edu/etd/3083

This Dissertation is brought to you for free and open access by University of Memphis Digital Commons. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of University of Memphis Digital Commons. For more information, please contact khggerty@memphis.edu.

CONTINUOUS ESTIMATION OF SMOKING LAPSE RISK FROM NOISY WRIST SENSOR DATA USING SPARSE AND POSITIVE-ONLY LABELS

by

Md Azim Ullah

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Computer Science

The University of Memphis

May 2023

© Copyright 2023 Md Azim Ullah

Partial rights reserved

ACKNOWLEDGMENTS

I sincerely thank my advisor, Dr. Santosh Kumar, who has inspired, motivated, and supported me with incredible patience while pursuing my doctoral research. It would be the most outstanding achievement of my life to have succeeded in learning from his high standards in research, interpersonal communication, and leadership skills. I am incredibly fortunate to have Dr. Kumar as my advisor, who has been a father figure to me in a foreign land and has patiently guided me through countless failures and frustrations in the quest for my dreams and aspirations.

I would also like to thank my committee members, Dr. Nirman Kumar, Dr. Myounggyu Won, and Dr. Xiaolei Huang, for agreeing to serve as members of my doctoral committee. I am immensely fortunate to have them as my committee members and cherish the opportunity to learn from their invaluable suggestions and comments.

I would also like to thank my lab members and colleagues, Dr. Soujanya Chatterjee, Dr. Nazir Saleheen, Dr. Timothy Hnat, Dr. Nasir Ali, Sameer Neupane, Mithun Saha, Rabin Banjade, Sayma Akther, Hosneara Ahmed, Shiplu Hawlader, Dr. Monowar Hossain, Dr. Rumanna Bari, Nusrat Nasrin, Dr. Anandatirtha Nandugudi, Shahin Alan Samiei, Joseph Biggers, Cheryl Hayes, and Lyndsey Rush for their encouragement and support. I sincerely thank my collaborators Dr. Benjamin Marlin, Dr. Emre Ertin, Dr. James Rehg, Dr. Mustafa al' Absi, Dr. Deniz Ones, Dr. Cho Lam, and Dr. David W. Wetter. I want to thank Dr. Benjamin Marlin for guiding me in a big part of my Ph.D. journey. I would also take this opportunity to thank all the smoking cessation research study coordinators at the University of Minnesota, Memphis, Utah, and Rice University, including Rebecca Stoffel, Michelle Chen, Kristi Parker, Jeffrey Ramirez, and Andy Leung, for their help in conducting the user studies in the natural environment which have been an integral part of my dissertation. I also wish to thank Shahin Samiei, Dr. Timothy Hnat, and Dr. Syed Monowar Hossain from MD2K Center

of Excellence at the University of Memphis for their contributions to data collection and/or software used for data collection.

I want to acknowledge my parents, Md Ahsan Ullah and Yasmin Begum, for all their sacrifices and love for me. Nothing would have been possible without their support, affection, and love. I would also like to mention my sister Shohely Sultana who has inspired me to go on in the most challenging times. I want to convey my gratitude to my dearest and lovely wife, Tamanna Ferdous. Her love, sacrifices, and support have kept me on this journey, and we have been lucky enough to be parents of our wonderful daughter, Anumegha. I want to express my fatherly love to her, born during my Ph.D. journey. Finally, I thank my friends in Memphis with a special mention to Dr. Maminur Islam. Over the years, Memphis has become my home. I am grateful to my friends in Memphis beyond all measures.

This research was supported in part by the National Institutes of Health (NIH) under awards P41EB028242, R01CA224537, R01MD010362, R01CA190329, U01CA229437, and by the National Science Foundation (NSF) under awards ACI-1640813, CNS-1823221, CNS-1705135, and CNS-1822935.

ABSTRACT

Ullah, Md Azim. Ph.D. The University of Memphis. May , 2023. Continuous Estimation of Smoking Lapse Risk from Noisy Wrist Sensor Data using Sparse and Positive-Only Labels. Major Professor: Dr. Santosh Kumar.

Estimating the imminent risk of adverse health behaviors provides opportunities for developing effective behavioral intervention mechanisms to prevent the occurrence of the target behavior. One of the key goals is to find opportune moments for intervention by passively detecting the rising risk of an imminent adverse behavior. Significant progress in mobile health research and the ability to continuously sense internal and external states of individual health and behavior has paved the way for detecting diverse risk factors from mobile sensor data. The next frontier in this research is to account for the combined effects of these risk factors to produce a composite risk score of adverse behaviors using wearable sensors convenient for daily use.

Developing a machine learning-based model for assessing the risk of smoking lapse in the natural environment faces significant outstanding challenges requiring the development of novel and unique methodologies for each of them. The first challenge is coming up with an accurate representation of noisy and incomplete sensor data to encode the present and historical influence of behavioral cues, mental states, and the interactions of individuals with their ever-changing environment. The next noteworthy challenge is the absence of confirmed negative labels of low-risk states and adequate precise annotations of high-risk states. Finally, the model should work on convenient wearable devices to facilitate widespread adoption in research and practice.

In this dissertation, we develop methods that account for the multi-faceted nature of smoking lapse behavior to train and evaluate a machine learning model capable of estimating composite risk scores in the natural environment. We first develop *mRisk*, which combines the effects of various mHealth biomarkers such as stress, physical activity, and location history in producing the risk of smoking lapse using sequential deep neural networks. We propose an event-based encoding of sensor data to reduce the

effect of noises and then present an approach to efficiently model the historical influence of recent and past sensor-derived contexts on the likelihood of smoking lapse. To circumvent the lack of confirmed negative labels (i.e., annotated low-risk moments) and only a few positive labels (i.e., sensor-based detection of smoking lapse corroborated by self-reports), we propose a new loss function to accurately optimize the models.

We build the *mRisk* models using biomarker (stress, physical activity) streams derived from chest-worn sensors. Adapting the models to work with less invasive and more convenient wrist-based sensors requires adapting the biomarker detection models to work with wrist-worn sensor data. To that end, we develop robust stress and activity inference methodologies from noisy wrist-sensor data. We first propose *CQP*, which quantifies wrist-sensor collected PPG data quality. Next, we show that integrating *CQP* within the inference pipeline improves accuracy-yield trade-offs associated with stress detection from wrist-worn PPG sensors in the natural environment. *mRisk* also requires sensor-based precise detection of smoking events and confirmation through self-reports to extract positive labels. Hence, we develop *rSmoke*, an orientation-invariant smoking detection model that is robust to the variations in sensor data resulting from orientation switches in the field.

We train the proposed *mRisk* risk estimation models using the wrist-based inferences of lapse risk factors. To evaluate the utility of the risk models, we simulate the delivery of intelligent smoking interventions to at-risk participants as informed by the composite risk scores. Our results demonstrate the envisaged impact of machine learning-based models operating on wrist-worn wearable sensor data to output continuous smoking lapse risk scores. The novel methodologies we propose throughout this dissertation help instigate a new frontier in smoking research that can potentially improve the smoking abstinence rate in participants willing to quit.

TABLE OF CONTENTS

Contents **Pages**

List of Figures x

List of Tables xiii

1 Introduction **1**
- 1.1 Overview and Motivation 1
- 1.2 Problem Statement 5
- 1.3 Challenges & Approach 7
- 1.4 Contributions 12
  - 1.4.1 Encoding the Decaying Historical Influence of Events 12
  - 1.4.2 Rare Positive Loss Function to Learn from Sparse Positive-only Labels 13
  - 1.4.3 Integrating Data Quality to Improve Inferences from Noisy Sensor Data 14
  - 1.4.4 Making Smoking Detection Robust to Orientation Switches 15
- 1.5 Dissertation Outline 16

2 Literature Review **19**
- 2.1 Introduction 19
- 2.2 Predicting the Risk of Adverse Events 20
- 2.3 Identifying the Risk Factors of Smoking Lapse Behavior 22
- 2.4 Detection of Risk Factors Using Mobile Sensors 23
- 2.5 Continuous Inference of Stress and Activity Using Wrist-worn Sensors 24
- 2.6 Smoking Event Detection Using Wrist-only Sensors 27
- 2.7 Orientation-Invariant Approaches to Dealing with Inertial Sensor Data 28
- 2.8 Learning from Sparse Positive-only Labels 31
- 2.9 Chapter Summary 32

3 mRisk: Continuous Risk Estimation for Smoking Lapse from Noisy Sensor Data with Incomplete and Positive-Only Labels **33**
- 3.1 Introduction 33
- 3.2 Smoking Cessation Study and Data Description 36
  - 3.2.1 Smoking Cessation Research 36
  - 3.2.2 Participants 37
  - 3.2.3 Study Protocol 37
  - 3.2.4 Wearable Sensors and Smartphone 38
  - 3.2.5 Determining the Smoking Lapse Time 38
  - 3.2.6 Data Selected for Modeling 40
- 3.3 Problem Setup and Formulation 40
  - 3.3.1 Problem Formulation 40
- 3.4 Robust Computation of Psychological, Behavioral & Environmental Context 42
  - 3.4.1 Robust Representation of the Current Context 43
  - 3.4.2 Encapsulating History via Events-of-Influence 44

Stress Events 45
Activity Events 46
Visitations to Smoking Spots 46
3.5 *mRisk*: Modeling Imminent risk of lapse 47
3.5.1 Deep Model with Recent Event Summarization (DRES) 47
Events-of-Influence Representation using Features 47
Feature Set 48
DRES Model Architecture 48
3.5.2 Deep Model with Decaying Historical Influence (DDHI) 49
Modeling Rationale 50
Decay-aware Temporal Encoding of Heterogeneous Events 51
Phenotyping Participants for Parameter Estimation 53
DDHI Model Architecture 54
3.6 Learning From Sparse & Positive only Labels 55
3.6.1 Positive Unlabeled (*PU*) Learning 55
3.6.2 Rare-Positive (*RP*) Loss Function 56
Design of the RP Loss Function 56
The Loss Function 59
3.7 Optimization, Evaluation, and Explanations of *mRisk* Model Choices 60
3.7.1 Loss Function Optimization and Evaluation 60
3.7.2 Evaluating *mRisk* Model Choices by Their Risk Characteristics 62
Results 63
3.7.3 Evaluating *mRisk* Model Choices via Simulated Delivery of Risk-Triggered Interventions 63
Evaluation Metric 64
Experiment Setup 65
Results 66
3.7.4 Evaluating *mRisk* Model Performance on Training-Independent EMA Labels 67
3.7.5 Rise/Fall in Risk Levels Produced by *mRisk* Before/After Lapse Moments 69
3.7.6 Understanding the Role of Context in Estimating Lapse Risk via Model Explanations 70
Observations from Global Feature Importance 71
3.8 Discussion, Limitations, and Future Works 72
3.8.1 Key New Insights 73
3.8.2 Limitations and Future Works 73
3.9 Chapter Summary 75

4 Robust Inference of Human States from More Convenient, but Noisier Wrist Sensor Data 77
4.1 Introduction 77
Adapting Inference Models from Chest to Wrist 79
4.2 Activity Recognition From Wrist-worn Accelerometry 81
4.3 Robust Stress Inference from Wrist-worn PPG 82

4.4 Our Approach & Key Contributions 84
4.5 Related Work on PPG Signal Quality Assessment in Light of Physiological Event Inference 86
4.5.1 Quality assessment of PPG Data 86
4.5.2 Using Signal Quality to Restore or Repair PPG Data 87
4.5.3 Signal Quality for Physiological Event Inference 88
4.6 Datasets 89
4.6.1 Devices 89
4.6.2 Study Protocols 90
Lab Study Protocol 91
Field study Protocol 91
4.6.3 Data Collected 92
4.7 CQP Data Quality Model 93
4.7.1 Windowing Data for Quality Assessment 93
4.7.2 Preprocessing PPG Data 94
4.7.3 Identifying and Isolating Irrecoverable PPG Segments 95
4.7.4 PPG Signal Quality Features 96
4.7.5 PPG Data Quality Labeling 97
4.7.6 The CQP-5 Model for Assessment of PPG Data Quality Over Five-Second Windows 99
4.7.7 Five-Second PPG Data Quality Classification Experiments 100
4.7.8 Five-Second Instantaneous Heart Rate Estimation Experiments 103
4.7.9 The CQP-60 Index for Assessment of PPG Data Quality Over Sixty-Second Windows 104
4.7.10 60-Second PPG Feature Computation Experiments 105
4.8 Deep Integration of the CQP Model to Improve the Robustness of PPG-Based Stress Inference 108
4.8.1 Stress Inference from ECG Data 108
4.8.2 PPG Data Cleaning and Computation of Inter-beat Intervals 109
4.8.3 Quality-Integrated PPG Feature Computation 109
4.8.4 PPG Feature Normalization to Account for Between-Person Variability 110
4.8.5 PPG Stress Model Training 111
4.8.6 Experiments on Accuracy vs. Yield of Stress Inference in the Lab 112
4.8.7 Experiments on Accuracy vs. Yield of Stress Inference in the Field 114
4.9 Limitations and Future Works 116
4.10 Chapter Summary 117

5 rSmoke: Orientation-Invariant Detection of Smoking Events from Wrist-worn Inertial Sensors 118
5.1 Introduction 118
5.2 Robustness Challenges to Smoking Detection using Wearable Wrist-worn Sensors in the Field 120
Variability in Sensor Configurations 121
Variability in Axes Orientation Owing to Sensor Placement 121

Lack of Sufficient Training Data from Natural Environment 122
5.3 Prior Works on Smoking Detection and Our Contributions 122
5.4 Dataset Description 125
5.4.1 Training Data for Smoking Detection 125
5.4.2 Smoking Data from Field With Original Sensor Mount 126
5.4.3 Smoking Data from Field With Switched Sensor Mount 126
5.5 Inertial Sensor Mount Identification and Axis Alignment 126
5.5.1 Distribution of Vertical Axis during Walking 128
5.5.2 Key Idea: Distinguishing between Lateral and Perpendicular Axes Distribution during Walking 129
Distribution of Lateral Axis 129
Distribution of Perpendicular Axis 130
5.5.3 Identification of 3-Axis Inertial Sensor Configuration 131
5.5.4 Investigating the possibility of Exact Alignment of Individual Accelerometer Axes 132
5.6 *rSmoke*: Smoking Episode Detection 133
5.6.1 Data Preprocessing 133
5.6.2 Smoking Puff Detection 134
Building Upon Prior Works By Identifying the Limitations of Existing Puff Detection Model 134
Candidate Segment Generation and Selection 136
Orientation Invariant Features from Candidate Puffs 137
Puff Detection Model 138
5.6.3 Smoking Episode Construction from Noisy Detected Puffs and Event Modelling 141
Constructing Candidate Smoking Episodes from Detected Smoking Puffs 141
Excluding Non-Smoking Episodes Based on Duration and Count of Detected Puffs 142
Representing Candidate Smoking Episodes for Learning 143
Smoking Event Detection Model: 145
5.7 Performance on Detecting EMA-Reported Smoking Events 146
5.7.1 Self-Reporting Smoking Occurrence using EMAs 146
5.7.2 Results 147
5.8 Limitations and Future Works 149
5.9 Chapter Summary 150

6 Continuous Assessment of Smoking Lapse Risk From Wrist-worn Sensors 151
6.1 Introduction 151
6.2 Smoking Cessation Research Study with both Wrist and Chest Sensors 153
6.2.1 Study Participants Recruitment and Protocol 153
6.2.2 Wearable Sensor Suites 154
6.2.3 Data Volume 154
6.3 Wrist-based Smoking Lapse Risk Estimation 155

6.3.1 Data Processing 155
6.3.2 Risk Estimation Models 157
6.4 Intervention Delivery Informed by Risk Episodes 159
6.4.1 Finding Opportune Moments for Smoking Interventions 159
6.5 Evaluating the Effectiveness of Risk Peak Triggered Simulated Interventions 161
6.6 Discussions, Limitations and Future Works 164
6.7 Chapter Summary 165

7 Conclusion and Future Directions 167
7.1 Summary and Key Contributions 167
7.2 Future Research Directions 169

LIST OF FIGURES

Figures **Pages**

3.1 Internal state and external cues from an observation window and prior to it are used to estimate the risk of a smoking lapse during the prediction window. The intervention window between the observation and prediction windows gives an opportunity to deliver an intervention. 41

3.2 (a) Sensors and extracted events used for model development (b) A stress stream with a stress event 45

3.3 Overall architecture of the Deep Model with Recent Event Summarization (DRES) 49

3.4 Architecture of the Deep Model with Decaying Historical Influence (DDHI) that uses an explicit model of decaying influence of past events that are expected to wane over time 54

3.5 P and R values when using different values of ϵ in the *RP* loss function, compared with that from using Triplet loss 60

3.6 Evaluating *mRisk* model choices on *PU-labeled* data 62

3.7 (a) Shows the EMA items corresponding to smoking report by individuals, (b) Intervention Hit Rate at 5.5 int. per day when considering a certain duration of EMA response as positive lapse 67

3.8 Lapse Likelihood produced by the *DDHI* model with lapse, intervention and EMA report times shown with vertical lines. We only include those EMAs in which the participants confirmed that the last time they smoked was 0-2 hours ago. 69

3.9 Global Feature Importance showing top 10 features for *DRES* model using *Deep SHAP* 71

4.1 Confusion Matrix of Activity Classification in WISDM dataset 81

4.2 Lab study protocol and devices used for data collection. The chestband consists of ECG, respiration, and accelerometers. The wrist device consists of 3 channels of PPG, accelerometers, and gyroscope. 89

4.3 Figures a-g show annotation of single channel PPG windows with respective power spectral density plot on top. The vertical lines in the top subplot show the heart rate frequency range. Spectral peaks inside and outside this range are colored separately. Figures (a) and (e) show acceptable segments, (b) and (f) show undecidable segments, (c) and (g) show unacceptable segments and (d) shows irrecoverable segments Figure (h) shows the intersection of different classes assigned to windows through a Venn Diagram. Irrecoverable subclass is shown inside a gap within the Unacceptable class. 98

4.4 Signal Quality Classification Results.Figure (a) shows train and test ROC for CQP model, test ROC for skewness and inertial motion using decision tree model. Figure (b) shows train and test confusion matrices for two classes: Undecidable/unacceptable, and acceptable. Training confusion matrix is obtained after 10 fold-stratified cross validation applied to training data. Figure (c) shows the error distribution in heart rate estimation in different bins of signal quality likelihood for each 5-second segment of PPG data. 103

4.5 Yield-Accuracy Trade-off of Quality Aware vs. not aware stress modelling. Figure(a) shows the Confusion Matrices for ECG stress model and PPG Based Stress Model from left wrist. Figure(b) shows the Leave one subject out cross validation scores of PPG stress model developed on data from left wrist in the lab as a function of increasing minimum thresholds on Mean-CQP-60. Figure(c) shows data yield vs correlation with ECG stress model of both quality integrated and not integrated PPG based stress model in lab. Green represents quality integrated model whereas red represents quality not integrated model. X-axis in both (b) & (c) shows the mean number of minutes available in field for corresponding thresholds on Mean-CQP-60 113

5.1 Figures showing sensor mounting and axes orientation variability.(Figure (b) taken from [1]) 121

5.2 Figures showing (a) opposing orientation of the perpendicular axis between left and right wrist, (b) three different walking moments showing values of lateral and perpendicular axis in the right wrist (human figurine copied from [2]), (c) Distribution of the accelerometer axes values during walking in our studies 127

5.3 Figures showing inertial wrist sensor data of both wrists during smoking episodes 135

5.4 Histogram of the Duration of All Labelled Smoking Puffs 137

5.5 Performance of Smoking Puff Detection Model 139

5.6 Distribution of the Inter Smoking Puff duration and count in labeled data 141

5.7 (a) Distribution of the Duration of Candidate Smoking Episodes, (b) Distribution of the count of actual puffs within candidate smoking episodes, (c) Distribution of detected and smoking puff counts within smoking episodes 143

5.8 Confusion Matrix for Smoking Event Detection 145

6.1 Examples of Lapse Risk Episodes in daily Smoking Lapse Risk Scores 158

6.2 Wrist-derived *mRisk* Lapse likelihood averaged across all the lapse moments 163

LIST OF TABLES

Tables **Pages**

3.1 Intervention Hit Rate at Different Frequencies of Intervention for Different Models 66

3.2 Intervention Hit Rates obtained from *DDHI* model with different number of phenotypes 66

4.1 Description of data from the Lab & Field Studies 92

4.2 Inter-rater distribution statistics for calculation of kappa statistic κ 99

4.3 PPG Signal Quality Model Performance on 5-second Segments 102

4.4 Correlation of Minute level HRV features computed from PPG in Lab 106

4.5 Correlation of Minute level HRV features computed from PPG in Field 107

4.6 Yield breakdown on field data using Mean-CQP-60 ≥ 0.2 112

4.7 Correlation Between ECG and PPG-based Stress Models and Yield of total data in Field with and without Quality-Integration 115

5.1 Performance of *rSmoke* model from smoking self-reports 147

6.1 Intervention Hit Rate at different daily frequencies of intervention using wrist sensors 161

6.2 Intervention Hit Rate of the DDHI model at different daily frequencies of intervention using different sensing modalities 162

Chapter 1

Introduction

1.1 Overview and Motivation

With more than 8 million deaths per year globally from tobacco use, smoking remains a leading cause of morbidity and mortality across the world [3]. Considering the current prevalence rates across the world population, researchers estimate that smoking can cause upwards of 1 billion preventable deaths in the twenty-first century [4, 5]. Hence, the success of smoking cessation programs (both behavioral and pharmacological) is of paramount importance. Although most adult smokers (68.0%) want to quit and more than half (55.1%) have made a quit attempt, the success rate remains drastically low, with fewer than one in ten (7.5%) adult cigarette smokers succeeding in quitting each year [6]. Thus, comprehensive multi-disciplinary approaches toward effective management of factors affecting smoking cessation success remain a crucial area of research spanning many academic fields. In recent times, emerging behavioral modification programs that deliver just-in-time smoking interventions (JITAIs) offer great promise with the ability to pinpoint moments of vulnerability and fight off smoking relapse [7]. Increasingly powerful mobile health (mHealth) sensing technologies underpin this adaptive intervention design strategy [8].

Advances in mobile health sensing research coupled with the increasing availability of smartphones, wearables, and the concurrent rise of cloud and edge computing have allowed for continuous, remote monitoring of individuals' health and wellness factors. Sensors embedded in the edge computing devices such as GPS, inertial motion sensors (Accelerometer and Gyroscope), and physiological sensors (Electrocardiogram, Respiratory Inductive Plethysmograph, and Photoplethysmogram) have enabled the collection of an enormous volume of sensor data. These data streams often contain unique manifestations of individual behaviors, and researchers have analyzed them for detecting behavioral events, mental states, and others using machine

learning models. Examples include physical activity [9], smoking [10], brushing [1], sleep [11], drinking [12], conversation [13], and eating [14]; health states and conditions such as mental stress [15], depression [16, 17, 18], epilepsy seizures [19], asthma [20], and glaucoma [21] as well as contextual factors such as social interactions [13] and mobility [18]. These machine learning-based detection models infer the events from unique signatures within the sensor signals, which manifest from an occurrence of the events themselves. The ability to infer the aforementioned behaviors, individual contexts, and health status of individuals in the natural environment allows researchers to represent the current context and status of participants effectively. Thus, designing and developing effective just-in-time adaptive intervention strategies for behavioral modification and treatment delivery becomes possible.

In smoking cessation programs, individuals attempt to quit smoking voluntarily and begin their abstinence period. We term "smoking lapse" as reverting to smoking during the abstinence period. The first lapse represents a transition from abstinence to smoking with the majority of all lapses eventuating in a full relapse [22, 23]. We aim to predict the risk of any impending lapse behavior. The goal is first to identify the at-risk moments and this will facilitate intervention design and delivery at different points in time. In contrast to singular events mentioned beforehand, smoking lapse is a multi-faceted behavior depending on various factors. Such as the inability to inhibit acting to intense withdrawal, stress arousal, urges/cravings, or failure to self-regulate or self-control under conducive environmental or situational cues [24]. And, since the "lapse" phenomenon has not occurred yet, the signatures rooted within sensor signals are not pronounced. Thus, outputting a risk score of smoking lapse using mHealth sensors first requires identifying the dynamic risk factors affecting lapse behavior and employing measures that can detect the occurrence of these factors in the natural environment.

Research [25, 23, 8] have brought into light both the internal and external factors which influence the onset of smoking lapse resulting in full smoking relapse.

First, negative affect has been consistently associated with lapse behavior acting as an internal trigger [26, 27, 28, 29, 23]. Positive affect situations where individuals exhibit emotionally positive situations can also precede lapse events [29, 23, 27]. Continuous estimation of stress [15], craving [30] using mobile sensors in the field provide us the opportunity to passively detect the internal triggers related to positive and negative affect situations. Second, exposure to external stimuli such as proximity to a bar or seeing others smoke increases the chances of a smoking lapse [31]. Smoking opportunity context [32] detects the exposure to smoking spots and represents the situational cues using GPS sensors. Detecting these risk factors in isolation and triggering interventions based on the occurrence of any of these predetermined events do not offer a comprehensive approach to estimating the risk of smoking lapse. We must consider the combined effects of both the internal and external stimuli, compose both triggers together using mobile sensors, and represent them accordingly to produce a single composite risk score.

The mobile sensors we employ must be convenient to wear in our daily lives and, simultaneously, be able to sense an individual's context accurately and without significant gaps in sensing. Wrist-worn wearable sensors are the most suitable option as they are non-invasive, easy to use, and widely available. The models published in the literature to estimate the risk factors usually employ more accurate chest-worn sensors. For example, physiological stress detection models use chest-worn ECG for inferring the heartbeat timings [15]. Acclimating the stress detection model to work with wrist-worn PPG sensors requires careful consideration of the initially employed methodologies and adapting them to fit the new sensing context. A substantial challenge thus stems from accurately estimating the internal and external triggers using wrist-worn wearables. Wrist sensors are more susceptible to noise owing to their peripheral placement. Uncertainties due to improper sensor placement, transient wrist motion, and other factors introduce important data quality considerations in subsequent inferences.

Works on predicting the risk of adverse events include risk prediction of adverse clinical outcomes such as ICU admission [33, 34], mortality [35, 36, 37, 38], and disease diagnosis [39, 40, 41, 42, 43] as well as adverse events related to public safety and disasters such as fire [44, 45, 46], accidents [47, 48, 49], flood [50], and wildfire [51, 52]. Our work is unique since we aim to output the risk of adverse behavior instead of concrete events. Smoking lapse behavior is not wholly and precisely observed as opposed to an event such as clinical death, fire, or others. Developing a machine learning-based model for outputting the risk of smoking lapse requires ground truth labels of high-risk (positive) and low-risk (negative) instances. First, we lack any negative labels of low-risk moments. Since our observations of the participant's actions are limited in scope, and taking into consideration the coping ability even when there is an urge to smoke, we can not confidently locate low-risk moments within the abstinence period. On the other hand, if ascertained correctly, precise lapse moments act as the source of ground truth labels for high-risk moments. However, in contrast to the previous adverse events, obtaining the precise timing of the lapse is challenging. Earlier research used retrospective recalls of smoking lapse situations from participants' memory (sometimes months old) [28, 53] to construct a coarse time frame of lapse behavior. To circumvent the limitations owing from autobiographical memory-based reconstruction, *Shiffman et al.* proposed the use of Ecological Momentary Assessments (EMAs) [54] for collecting data on lapse and temptation antecedents close to real-time and in participants' natural environment [23]. Even with EMAs, abstinent individuals can only provide a retrospective self-report of lapse situations, lacking adequate precision for exact annotations. We depend on sensor-based detection of smoking events to obtain the precise timing of smoking.

Existing smoking detection models employ one or a combination of sensing modalities to detect smoking behavior [55, 56]. These studies typically involve data collection using a constrained lab setup with participants wearing one or more sensors

placed in a reference position [57, 58, 59, 60, 61, 55, 56, 58, 10, 62]. The constrained lab environment and the fixed reference position of the inertial sensors limits the utility of the developed models in the field. Outputting the risk of an adverse behavior like smoking lapse using convenient wrist-worn wearable sensors requires a working smoking detection model from wrist-worn inertial motion sensors alone. Variability in sensor configurations, sensor placement resulting in variability in axes direction, lack of sufficient training data, and difficulty in collecting reliable ground truths are some of the many challenges facing the development of a robust smoking detection model from wrist sensors in the field. Any developed model must be robust to the multitude of circumstances that affect the accurate detection of smoking events using wrist-based wearables in the natural environment.

As participants may not wear the sensors at the time of lapse, and the sensor-based detection models can miss the lapse events, we can not capture the timing of each smoking lapse. Therefore, we only have sparse positive only labels for training the risk estimation models. Using these sparse positive-only labels obtained from the precise detection of smoking events and corroboration through self-reports, we need to train and optimize our models to recognize the high-risk moments. We must devise a novel methodology to represent participants' current and past contexts derived from noisy sensor data. The model must learn within the constraints of incomplete and sparse labels present and output risk scores that demonstrate the ability to inform the design and delivery of just-in-time smoking interventions.

1.2 Problem Statement

We aim to predict the imminent risk of a smoking lapse using mobile and wearable sensing in individuals' natural environments. We seek to train a machine learning-based model capable of continuously producing the risk of smoking lapse using noisy sensor data from wrist-worn wearables. Solving these problems concerns

overcoming several challenges. We briefly describe these challenges here and elaborate on them in Section 1.3

The first challenge involves estimating individuals' context from noisy sensor data in the natural environment and using them to learn a lapse risk prediction model. Characterizing individuals' current and historical context using mobile sensor-derived data streams requires developing novel time-series representation methodologies that are generalizable across different data sources and allow efficient and accurate model learning. The models selected also need to be appropriate for learning from multidimensional, multi-faceted sources of information.

The second challenge involves learning the risk prediction model using incomplete and imprecise ground truth labels. We need concrete and trustworthy labeled data instances to train our model. We lack negative labels and have only a few positively labeled samples. Training the risk prediction model in this scenario requires thoroughly assessing the available labeled and unlabeled data. We must employ novel methodologies for learning the dynamics of the smoking lapse phenomenon using sparse positive-only labeled instances.

The third challenge involves the nature of mobile health sensors employed to passively sense the participants' physiological, behavioral, and environmental context. The sensors must be convenient to wear for participants in their daily lives. They should be less invasive in form and functionality and allow for continuous sensing with minimal demands on the participants' behalf. Ensuring that the risk prediction model works with data collected from conveniently worn sensors is essential to the practical utility of our developed models and methodologies.

In this dissertation, we propose methods to address the above challenges. We demonstrate the feasibility of our methods and models with rigorous evaluation of developed processes using data from real-life smoking cessation field studies.

1.3 Challenges & Approach

This section elaborates on the specific challenges in achieving our goal. We also outline the approach we develop to deal with those challenges.

- **Representing Individual Context using Noisy Sensor Derived Observations:** Observations of participants' context and health states using mobile sensors yield numerous multidimensional streams differing in scale, frequency, alignment, and other attributes. Several data quality factors, such as improper sensor wearing, dynamic motion, software glitches, and participant non-compliance, introduce significant noise within these data sources. The foremost step in using the mHealth sensor data is transforming the noise-corrupted data sources into representations of individual health, wellness states, behaviors, and actions. With substantial diversity in the individual and collective representation (e.g., frequency, duration, type, etc.), these intermediate streams also suffer from rapid variability, noise, and discontinuity. Using them to predict the risk of smoking lapse requires selecting the most suitable model capable of extracting meaning and learning from the dynamic interactions of these noisy contexts and their representations. We must explore novel ways of accurately representing participants' context that fits our data, selected model, and problem setup. Moreover, our model should be capable of taking in multivariate input sources of contextual information from participants' past and present and produce a composite risk of smoking lapse.

  Our Approach: From the collected sensor data in the natural environment, we compute the dynamic risk factors representing psychological (e.g., stress), behavioral (e.g., activity), and environmental (e.g., proximity to a smoking spot) contexts using state-of-the-art methods from the literature. We call these "continuous inference streams" since we infer them continuously using trained machine learning-based models whenever data is available. The continuous

inference streams are stable, intermediate, and lower frequency representations of raw sensor data. We then encode the inference streams into events. Events are sparse locations within the constant inference streams that contain information about the influence of the underlying risk factor impacting participants' lives. We term them as 'events-of-influence' streams. We compute the homogeneous statistical representations of continuous inference streams and events-of-influence time series and train deep learning models to output the risk of smoking lapse. The deep neural network-based models are used as universal approximators of the underlying risk dynamics and can exploit temporal and spatial patterns in our data. Finally, we explore approaches to succinctly capture the historical influence of recent and past events (i.e., substantial change in any context) to make deep learning models efficient. First, we summarize the influence of recent and past events via new features. Second, we explicitly encode the impact of current and past events as an exponentially decaying function over time. The proposed second approach, "Decay-aware Temporal Encoding of Heterogeneous Events," replaces the need for explicit feature engineering and provides a novel way of ingesting the events-of-influence data into deep machine neural network models for time-series-based applications in general.

- **Learning from sparse positive-only labels:** Training and optimizing machine learning-based models from mobile sensor data requires accurate annotations of ground truth behaviors, contexts, or actions. For our problem, a smoking lapse event indicates a high-risk moment in which the participant foregoes abstinence due to internal and external factors forcing him to revert to smoking. Only these lapse events allow us to impose annotations of risk levels onto the underlying sensor signals. We rely on sensor-based detection of smoking lapse events in the natural environment to detect smoking lapse behavior. Sensor-based detection of smoking events provides precise timings of the lapse so we can annotate the data

instances of those moments. However, detecting every smoking lapse behavior is impossible since participants do not always wear the sensors. Also, the models used to detect smoking suffer from imperfections leading to missed detections and false positives. Thus, even with sensor-based detection, we need confirmation through other means to provide accurate annotations with high confidence. All these factors culminate in us detecting only a subset of all possible lapse events. These confirmed lapses are few; thus, we only have sparse positive labels of high-risk moments. We also have no known method of obtaining labels for low-risk moments. Since we can have missed lapse events, and participants may not lapse even when the risk is high, we can not confidently say that participants were at low risk when smoking was not detected. Thus, a crucial challenge in our problem is the need to learn risk prediction models using sparse positive-only labels. Furthermore, since we aim to predict the risk of behavior that has not happened yet, the signatures within the sensor signals are not pronounced. Hence, it falls on the adopted models and methodologies to learn the ingrained patterns. Using only a handful of positive labels and without negative labels, learning a complex behavioral manifestation within the sensor data is an eccentric and challenging task.

Our Approach: We detect smoking lapse events using sensor-based smoking detection models. To further confirm the detected lapses, we use participant-provided self-reports through EMAs. Both strategies are incomplete in singularity since EMAs do not specify the timing of smoking lapse, and puffMarker can suffer from model and sensor imperfections. Hence, we only obtain a small number of confirmed lapse events, which are the source of positive labels of high-risk moments. For training our risk prediction model using these labeled instances, we embrace Positive Unlabeled Learning (PU Learning). We especially adopt the PU Bagging [63] model learning strategy when provided with positive

and unlabeled data instances. The PU Bagging approach to model learning combines a series of classifiers on datasets obtained by perturbing the initial training set through bootstrap re-sampling with replacement and blending these classifiers through aggregation. We train our deep neural network model according to the PU Bagging approach of ensemble model aggregation. We propose a novel loss function called "Rare Positive Loss" (RP Loss) to optimize our models and encode the data instances into a representational feature space. RP loss function is a metric learning loss function that guides the learning process so that our model accurately represents the positive class and also learns to extract other rare true positives from the unlabeled class.

- **Continuous and Robust Assessment using Convenient Sensors in the Natural Environment:** We aim to develop methodologies that allow for continuous estimation of lapse risk using noisy sensor data collected from the natural environment. We also need to employ convenient sensing and inference mechanisms to ease the burden on participants owing to continuous passive sensing. Numerous external and internal factors in the natural environment adversely impact the mobile sensing process and our ability to infer health or behavior states in the natural environment. These factors are often dynamic and come entangled with the original sensing medium. Our design of the risk prediction pipeline involves using several intermediate data streams representing the participants' physiological (e.g., stress), behavioral (e.g., activity), and environmental (e.g., location traces) contexts. To enable wrist-based inference of smoking lapse risk, we also need a detection model to accurately estimate smoking events in the natural environment. These intermediate inference streams are outputs of trained machine learning based-models and suffer from the impact of the data quality factors, sensing nuances, differences between employed sensors, and noise elements. These reduce the trustworthiness of the model output to end

users and ultimately affect the performance of the risk estimation model trained on top of them.

Our Approach: Our goal is to devise the risk prediction model using wrist-worn wearable sensors. Individuals are increasingly adopting wrist-worn wearables for analyzing their health and wellness states. These wearables are easy to wear and cause less burden on the individuals. They often are equipped with edge computing abilities, allowing for hosting simple ML models. We want to develop a smoking risk prediction model that takes inputs from commonly available and easy-to-wear wrist-based wearable sensors in the natural environment. The first essential prerequisite of realizing this goal is to adapt the intermediate inference streams from their native sensing mediums to wrist-based ones. Acclimating the physiological stress detection model to work with wrist-worn PPG sensors is a critical step in this direction. Wrist-based PPG data is susceptible to dynamic fluctuations of signal quality owing to increased peripheral motion and other nuances of reflective photoplethysmography. To this end, we propose *CQP*. *CQP* is a continuous data quality indicator that can be more deeply integrated into subsequent inferences to improve their robustness. *CQP* quantifies the signal quality of time-varying signals, informs the quality of inference, and allows for an improved accuracy-yield trade-off for stress inference using wrist-worn PPG in the field. Inferences using wrist-worn inertial motion sensors also suffer from problems related to axes orientation switching between different sensors in training and testing time and variability owing to changes in sensor placement. We use inertial sensors to infer participants' activity and smoking behavior in the field. For activity inference, we train our model using only the magnitude of the accelerometer sensor. Using the magnitude time series only, the developed model can be generalized across orientation differences in different devices and study setups. For robust detection of smoking events in the natural environment, we

propose *rSmoke*, an orientation invariant model of smoking detection from the 6-axis accelerometer and gyroscope sensors. *rSmoke* proposes a methodology to identify the individual axes configuration for inertial sensors and employs a robust feature computation and modeling framework for detecting smoking events from wrist-worn sensors in the field. The orientation invariant approach adopted in the *rSmoke* model shows robustness to changes in inertial sensor configuration and variability in axes orientation resulting from different sensor placements.

1.4 Contributions

In developing the approaches to deal with the challenges mentioned earlier, we innovate in several key areas. In this section, we enumerate our significant contributions in devising an end-to-end machine learning-based model capable of continuously outputting the imminent risk of smoking lapse using mobile sensors in the natural environment.

1.4.1 Encoding the Decaying Historical Influence of Events

In participants' natural environment, we use continuous assessments of stress, activity, and smoking opportunity to characterize participants' psychological, behavioral, and environmental contexts. We compute events-of-influence time series from these continuous inference streams similar to the approaches undertaken in [24]. Events allow us to observe and encode the historical dynamics of multiple different risk factors. However, encoding the historical context through sparse and episodic events-of-influence streams for training machine learning models is not straightforward. We employ two methods of representing the events-of-influence time series.

First, we represent the events data using features proposed in the literature [24] and propose to train a long-short-term-memory (LSTM) based deep neural network model(see Chapter 3.5.1). The choice of deep models aligns with their incredible popularity and success in encoding multivariate time-varying input data sources for applications such as forecasting, classification, and others. The deep neural network

articulates and brings into light the temporal, spatial, and interaction effects of underlying risk factors and produce a composite risk score. Moreover, a deep model allows us to employ customized data representations appropriate for our use case. Hence, for the second approach, we propose "Decay-aware Temporal Encoding of Heterogeneous Events," a novel encoding mechanism for the events-of-influence time series(see Chapter 3.5.2). The proposed encoding method traces the residual effects of recent and past events into the present using exponential decay functions. Therefore, it allows us to represent the historical contexts of participants. The encoding method fits appropriately with the temporal dimension of the LSTM model and removes the necessity of adopting explicit feature engineering approaches. We utilize participant phenotyping to estimate the degrees of freedom of our proposed encoding mechanism and train another deep neural network model (see Chapter 3.5.2). Results show that the second approach, the proposed event encoding scheme to capture the historical contexts in conjunction with the LSTM-based deep neural network, improves the accuracy of simulated just-in-time interventions compared to existing approaches.

1.4.2 Rare Positive Loss Function to Learn from Sparse Positive-only Labels

Our choice of deep neural network-based models allows for extracting higher-level patterns within the data and creating an accurate representation space of the training instances. We focus on optimizing this representational space with the available ground truth instances. The key to this approach is designing a loss objective appropriate for guiding the model learning process. Our proposed loss function, Rare Positive Loss (RP loss, see Chapter 3.6.2), is novel and works along two competing dimensions of model optimization within the constraints of positive unlabeled learning. First, we want to create a representational feature space in which positive data points create a tight cluster. We call this "Positive Class Dispersion." Ensuring ideal positive class dispersion alone is trivial for the model by merging all the input instances into a single point in the feature space. With the second condition, we constrain such development. Alongside the

positive class dispersion, we also want to ensure that the learned representation space of the positive class can only include a small portion of the unlabeled class. We call this the "Rarity of Unknown Positives Within Unlabeled Class." The overall loss function is a numerical combination of these two dimensions. Our results show that the proposed loss function outperforms more popular metric learning-based loss objectives unsuited to our cause. We obtain the best result when learning the model using the proposed event encoding methodology in conjunction with the RP loss function.

1.4.3 Integrating Data Quality to Improve Inferences from Noisy Sensor Data

Ongoing research in the mHealth domain seeks to infer continuous measures of physiological states and events from PPG sensors that are now integral components in smartwatches and activity trackers. This dissertation aims to develop a continuous smoking risk estimation model using convenient wrist-worn sensors. An essential prerequisite of fulfilling this goal is to enable robust inference of physiological stress levels using PPG data collected from wrist-worn sensors. However, persistent and dynamic noise elements in the PPG data make it particularly challenging and require novel, innovative methods. In Chapter 4, We proposed an approach to estimating PPG data quality using supervised learning and showed how the resulting continuous data quality indicator, CQP (see Chapter 4.7), can be more deeply integrated into subsequent inferences to improve accuracy-yield trade-offs for both the computation of individual features and complex high-level inferences such as stress. We devise a new approach toward auxiliary estimation and deep integration of signal quality metrics to inform and improve inference quality, accuracy, and robustness. Our results show that integrating signal quality levels within the inference mechanism enhances the accuracy and robustness of continuous inference from wrist-worn PPG sensor data. The resulting PPG-based stress model allows us to envisage the development of a smoking lapse risk estimation from conveniently-worn wrist-based wearables.

1.4.4 Making Smoking Detection Robust to Orientation Switches

We need a smoking detection model to accurately detect smoking events from wrist-worn inertial sensor data in the natural field environment to enable wrist-based estimation of smoking lapse risk. Smoking detection methods built on wrist-worn wearable sensors typically assume a fixed configuration of the inertial sensor on the wrists. They ensure this by constraining their data collection environment to the laboratory and enforcing compliance by using only one type of sensor with a fixed orientation. However, in the natural environment, these assumptions seldom hold. For example, the accelerometer x-axis in one sensor can be the y-axis in a different sensor. Even when the general axes are the same, the direction of the same individual axis can differ between sensors. Also, for the same sensor, the direction of inertial movement in inertial sensors can change dynamically owing to different sensor placements. Existing methods of smoking detection do not address these concerns and therefore lack the robustness to changes in sensor configurations and axes orientation. We propose *rSmoke*, an orientation-invariant approach to smoking detection that is robust to orientation switches in the field. Using the distribution of the sensor data in times of walking, we propose methods to identify the general direction of each axis in an inertial sensor. The proposed methodology of individual axis identification allows us to match the sensor configurations of different types of sensors. Once matched, we can align the configurations to a reference point. Once the general directions of the individual axes are known with respect to a reference direction, we proceed to identify the exact direction of the individual sensor axes using the same distribution. We propose methods to dynamically identify the exact direction and align the inertial accelerometer sensor's lateral axis (axis parallel to the direction of gravity). We also note the difficulty in finding the exact direction of the other two sensor axes. We base our feature computation and smoking puff candidate identification methodology on these findings to allow for the computation of robust features from inertial sensor data. Therefore, the

developed *rSmoke* model is resilient to changes in sensor orientation resulting from different sensor types or variability in sensor placement.

1.5 Dissertation Outline

In this dissertation, we propose end-to-end methodologies to develop a machine learning model capable of continuously outputting the imminent risk of smoking lapse using mobile sensors in the natural environment. We design and evaluate our methods with data from a real-life smoking cessation field study. This section outlines the chapters that bring our envisaged goals to fruition through the developed methods and processes.

Chapter 2 presents the background in form of the published literature related to our problem. We categorize the literature into distinct areas of research to showcase the different sub-problems we tackle for developing the smoking lapse risk prediction model from wrist-worn sensors.

Chapter 3 presents *mRisk*, a computation model for sensing the imminent risk of smoking lapse behavior using chest-worn mobile sensors. In this work, we propose an event-based encoding of sensor data to reduce the effect of noises and then present an approach to efficiently model the historical influence of recent and past sensor-derived contexts on the likelihood of smoking lapse. Next, to circumvent the lack of any confirmed negative labels (i.e., time periods with no high-risk moment), and only a few positive labels (i.e., detected adverse behavior), we propose a new loss function. We use 1,012 days of sensor and self-report data collected from 92 participants in a smoking cessation field study to train deep learning models to produce a continuous risk estimate for the likelihood of an impending smoking lapse. The risk dynamics produced by the model show that risk peaks an average of 44 minutes before a lapse. Simulations on field study data show that using our model can create intervention opportunities for 85% of lapses with 5.5 interventions per day.

Chapter 4 presents approaches to enable robust inference of stress and activity

from noisy wrist-worn sensor data. We first propose *CQP*, a machine learning based data quality indicator which informs the quality of inference from time-varying signals and is fitting for integration within the stress inference process. In this work, we propose an approach to estimating PPG data quality over short time windows using supervised learning and show how the resulting continuous data quality indicator, CQP, can be more deeply integrated into subsequent inferences to improve their robustness. Using 28,000+ labeled PPG segments, we show that CQP detects segments with acceptable data quality with 95% balanced accuracy compared to 80% using previous data quality measures. We integrate CQP inside the PPG-based stress detection pipeline and thoroughly evaluate our proposed methods for robust inference from wrist-worn sensor data. Using paired ECG and PPG data from both lab (n = 36) and field studies (n = 105), we show that integrating CQP into the PPG stress detection pipeline can significantly improve accuracy-yield trade-offs.

Chapter 5 presents methodologies for smoking detection from wrist-worn inertial sensor data. We identify the challenges associated with smoking detection in the natural environment using wrist-worn IMU sensors. These challenges include variability in sensor configurations, sensor placement resulting in differences in axes orientation, lack of sufficient training data from the natural field environment, and difficulty in the collection of reliable ground truths. We propose *rSmoke*, an orientation-invariant approach to first identifying the axes configuration for inertial sensors in the wild. *rSmoke* builds upon the existing works on smoking detection using mobile health sensors and proposes a robust feature computation and modeling framework for detecting smoking events from wrist-worn sensors in the wild. Our proposed methodology includes a novel smoking episode construction scheme that allows for the representation and identification of smoking episodes from noisy and spurious smoking puffs. We test our model in two smoking cessation research studies employing different inertial sensors. Our model provides superior performance compared to the existing works and shows robustness

when dealing with differences in sensor types, configurations, and orientation. We leverage the developed *rSmoke* model to more accurately detect the smoking lapse events of abstinent smokers in their post-quit period. This allows us to develop the smoking lapse risk estimation model using wrist-worn sensors presented in Chapter 6.

Chapter 6 combines previous chapters to develop continuous smoking risk estimation models from wrist-worn sensors. We use data from another smoking cessation research study where participants wore chest and wrist sensors in their natural environment. We first apply the continuous inference models presented in Chapter 4 to passively estimate dynamic risk factors using wrist-worn sensor data. Next, we apply the *rSmoke* smoking detection model from Chapter 5 to capture smoking lapse events representing ground truth high-risk moments in the post-quit smoking abstinence period. Finally, we train the smoking lapse risk estimation models proposed in Chapter 3. To simulate our model's ability to deliver intelligent smoking interventions to abstinent participants, we propose a new online intervention delivery mechanism based on risk episodes. The simulation results demonstrate that wrist-worn sensors perform similarly to chest-based ones in delivering just-in-time adaptive smoking interventions. The novel modeling ideas proposed in Chapter 3 also contributes towards improved performance in our study with different sensing modalities. Chapter 6 fulfills the ultimate objective of this dissertation by developing continuous smoking lapse risk estimation models from wrist-worn sensors.

Finally, Chapter 7 presents the concluding remarks and also discusses several exciting future research directions stemming from our dissertation.

Chapter 2

Literature Review

2.1 Introduction

The primary goal of this dissertation is continuous estimation of smoking lapse risk in the natural environment using convenient wrist-worn sensor data. Our path toward achieving this goal involves solving multiple distinct research problems. The approaches to these problems allow us to conceive and materialize a comprehensive set of methods for achieving our primary goal. This chapter provides an overview of the relevant literature on each problem we tackle. We build upon these prior works of distinct research areas to design and develop our solution.

In Chapter 3, we devise an end-to-end set of methods for estimating the risk of smoking lapse from chest-worn sensors. First, we delineate the works related to assessing the risk of different kinds of adverse events relevant to our problem of estimating the risk of smoking lapse. We contrast the uniqueness of smoking lapse behavior compared to other adverse events. Second, we identify the risk factors that impact smoking lapse behavior during abstinence. Next, we embrace established approaches to continuously estimate these risk factors from passive sensors in the natural environment. Using state-of-the-art models from literature, we employ continuous inference models to estimate stress, activity, and proximity to smoking spots using chest-worn ECG, Accelerometry, and smartphone-based GPS sensors, respectively. The estimates of these risk factors in the natural environment act as input to our proposed lapse risk estimation models. To adapt the developed models to work with convenient but noisier wrist-worn sensors, we developed methodologies for robust inference of stress and activity from wrist-worn sensor data. Chapter 4 presents a general set of methods for continuous inference of stress from wrist-worn PPG sensors in the natural environment. Differing from the existing approaches in the literature, we propose our method of first quantifying the quality of PPG signals and integrating it

within the stress inference process to improve accuracy, robustness, and feasibility for deployment in real-life scenarios. To enable wrist-based estimation of smoking lapse risk, we need a smoking detection model from wrist-sensors to detect smoking events from inertial sensor data in the field. Chapter 5 proposes *rSmoke* model, an orientation-invariant approach to first identifying the axes configuration for inertial sensors in the wild. Finally Chapter 6 trains the wrist based smoking lapse risk estimation models using inferences from the developed wrist-based models.

This chapter aims to construct an informed background of our approaches and methodologies. We hope to present the state of existing literature before delving into the details of our proposed methods.

2.2 Predicting the Risk of Adverse Events

Several works deal with predicting the risk of adverse events. We can further categorize these events into two subtypes - events related to clinical/health outcomes and events concerned with public safety and disasters. Works on predicting adverse clinical health outcomes include predicting mortality [35, 36, 37, 38], ICU admission[33, 34], disease diagnosis [39, 40, 41, 42, 43], clinical sepsis [64, 65], and others. Predicting risk of adverse public safety events include property fire hazards [44, 45, 46], flood [50], road accidents [47, 48, 49], and wildfire [51, 52], and others. Our problem is unique since we aim to output the risk of an adverse behavior instead of events mentioned so far. Adverse clinical or public safety events are fully observed in nature. Thus, researchers can obtain precise ground truth labels of positive and negative classes for training machine learning-based models to predict or forecast these events. A smoking lapse, on the contrary, can not be precisely observed with the same exactness or precision. To detect smoking lapses in the field, we depend on sensor-based detection of smoking events using hand-to-mouth gestures. We also need to confirm the detected smoking events from retrospective self-reports collected using EMAs close to the smoking time. A combination of sensor-based detection and

confirmation through EMA reports culminates in us obtaining only a subset of possible lapses. These detected lapses provide sparse positive labels of high-risk moments that we use to train our models. Also, the absence of the mentioned clinical and public safety-based adverse events in time indicates that the risk of those events was low, and these moments act as the source of negative labels in the field. Since participants do not always wear the sensors and may not lapse even when there is an urge to smoke (indicating a high-risk moment), we have no way of knowing the exact times when participants were at low risk of lapse. Thus, compared to the existing works on predicting the risk of adverse events, we only have a sparse set of positive labels for high-risk moments and no labels for low-risk moments. The influence of the input variables on the risk of smoking lapse is also not as clearly understood compared to adverse clinical or public safety events.

Our work in Chapter 3 closely relates to the dissertation authored in [24]. To our knowledge, [24] ideates the possibility of smoking lapse risk estimation from mobile sensor data. Similar to their approach, Chapter 3 uses continuous inference of risk factors in the natural environment using chest-worn mobile health sensors and builds a machine learning-based end-to-end model for lapse risk estimation. However, we extend and improve their approach in multiple different ways. First, We propose to train long-short-term-memory (LSTM) based deep neural network models to articulate and bring into light the temporal, spatial, and interaction effects of underlying risk factors and produce a composite risk score. The LSTM-based model chosen for our problem shows superior performance compared to the traditional models (e.g., Random Forest) employed by authors in [24]. Second, in line with the ability of deep models to ingest customized data representation, we propose a novel event encoding methodology to represent the historical context by accumulating the residual effects of past events. Finally, we propose a novel loss function to improve the performance of a positive-unlabeled learning-based smoking risk prediction model. Nevertheless, we

borrow from their work to identify the risk factors and use continuous inference models to estimate them from chest-worn sensors in Chapter 3. The following two sections - Section 2.3 and Section 2.4 delve into the published literature related to identifying and detecting risk factors impacting smoking lapse risk in the natural environment. We include them for completeness purposes to present our overall goal of estimating the risk of smoking lapse using convenient wrist-worn sensors.

2.3 Identifying the Risk Factors of Smoking Lapse Behavior

Research [25, 23, 8] have brought into light the factors which influence the onset of smoking lapse resulting in full smoking relapse. Negative affect have been consistently associated with lapse behavior acting as an internal trigger [26, 27, 28, 29, 23]. Positive affect situations where individuals exhibit emotionally positive situations can also precede lapse events [29, 23, 27]. Exposure to external stimuli such as proximity to a bar or seeing others smoke increases the chances of a lapse behavior [31]. The factors influencing lapse behavior can be categorized into two broad categories [25]. First is the internal precursors or the physiological/emotional states such as stress, urge, self-regulatory capacity, and others. For example, high-stress levels and low self-regulatory capacity may increase the risk of a smoking lapse [23, 66, 67]. The second category relates to environmental or social cues conducive to lapse behavior in the abstinence period. For example, increased availability of cigarettes in specific locations or seeing others smoke can significantly increase the chances of an imminent smoking lapse [23, 68]. The impact of these events on smoking lapse has been extensively studied using self-reports. The methods employed in these works to analyze self-reports are not readily transferable to sensor data. Nevertheless, these works educate us on the diverse nature of specific risk factors that can be used as inputs to predict the imminent risk of smoking lapse behavior.

2.4 Detection of Risk Factors Using Mobile Sensors

Continuous estimation of the imminent risk of a smoking lapse requires passive detection of the associated risk factors using mobile and wearable sensors. These risk factors represent the participants' internal and external context in the natural environment and allow our model to learn the dynamics of smoking lapse provided with the context variables. Advances in mobile and wearable sensing have enabled the development of computational models to detect individuals' health and wellness states. Published works on detecting stress [15] use wearable physiological sensors such as ECG and Respiration to detect stress levels by computing cardiac and heart rate variability features from individuals' heartbeat dynamics. Smartphone-based GPS sensor data have been used to detect trajectories of depression for individuals in their natural environment [18, 16]. Researchers have utilized the physiological and inertial sensing to continuously estimate the craving [30], alcohol consumption [69], and cocaine intake [70]. Human activity and gait recognition using inertial motion units (Accelerometer and Gyroscope) is an established field with decades of incremental improvements in research contributing to the integration and adoption by industry in commercial smartwatches and fitness trackers [71]. Smoking opportunity context [32] detects the exposure to smoking spots and represents the situational cues using GPS and Motion sensors. We leverage these works to employ continuous inference models of stress, activity and smoking opportunity contexts. Detecting these risk factors in isolation and triggering interventions based on the occurrence of any of these predetermined events do not offer a comprehensive approach to estimating the risk of smoking lapse. We must consider the combined effects of both the internal and external stimuli, compose both triggers together using mobile sensors, and represent them accordingly to produce a single composite risk score. We develop our smoking risk estimation model using various sensor suites deployed in real-life smoking cessation studies. We first build an end-to-end smoking risk estimation model using chest-worn

sensors. We compute the inference streams (stress, activity, and smoking opportunity contexts) using ECG and Accelerometry data from the AutoSense [72] chest-worn sensor suite and smartphone-collected GPS sensor data. We also rank the contributions of features representing different risk types using Explainable AI.

We propose to develop the smoking risk estimation model using convenient wrist-worn sensors. To achieve this goal, we must continually infer stress and activity levels using data from wrist-based sensors. Hence, we develop methods for robust continuous inference using wrist-worn sensor data. We focus on stress detection using wrist-worn PPG sensors since detecting stress from noisy PPG sensor data in the natural environment is challenging and requires significant adaptation of the existing approaches for stress detection from wearable ECG data.

2.5 Continuous Inference of Stress and Activity Using Wrist-worn Sensors

To achieve our stated goal of smoking risk estimation from wrist-worn sensor data, we must be able to infer stress and activity using wrist-worn PPG and inertial motion sensors in the field. We borrow existing literature to develop a deep neural network model of human activity recognition. Deep learning-based human activity recognition models have gained incredible popularity owing to their ability to learn from multi-dimensional wearable motion sensor data and accurately distinguish between complex human tasks [73]. Convolutional Neural Networks (CNN) offers an efficient deep model architecture for activity classification in the wild [74, 75]. We train a CNN-based activity recognition model for each 20-second data segment using publicly available WISDM dataset [76].

Inferring stress from PPG depends on accurately assessing cardiac, and heart rate variability (HRV) features from wrist-worn PPG signals. Due to their peripheral placement, dynamic wrist motion, and irregular attachment of wrist-worn sensors to the point of contact, PPG sensing in the natural environment suffers from various external noises and confounds. Therefore, robust stress inference from noisy wrist-worn PPG in

the natural environment first requires decision-making on the state of the input PPG signal to contain valid information about participants' heart rate dynamics. Substantial works exist on using PPG to estimate simple features of cardiac activity such as heart rate [77, 78, 79, 80, 81, 82, 83]. These past works show that PPG data can yield assessments of heart rate under controlled conditions that are as accurate as ECG and such methods are currently deployed in a variety of commercial off the shelf devices. A significant body of work also assesses more complex physiological and behavioral states and related activities based on cardiac features. This includes work on atrial fibrillation [84, 85], cocaine use [86, 87, 88], sleep apnea [89], and stress [90, 91, 92, 15, 93, 94]. However, these works have largely focused on deriving features from data provided by wearable ECG devices. Indeed, complex applications like arrhythmia detection and cocaine use detection that use ECG morphological structure cannot be easily adapted to PPG sensing as PPG data do not reflect the detailed morphological structure captured by ECG. In the case of stress specifically, early works focused on using ECG to identify cardiac features that correlate with elevated stress levels [95, 96, 97, 98, 99, 100]. These works led to the identification of ECG-derived HRV features as essential indicators of stress. These features are statistics of the inter-beat interval time series, making them amenable to assessment using wearable PPG devices. More recent studies have shown that HRV features can be accurately derived from PPG sensor data under controlled conditions [101]. The problem of stress detection from PPG data has subsequently been considered in several lab-based studies [102, 103, 104]. In [104], stress is detected in 20-second windows using peak-to-peak intervals from finger PPG and temperature recorded from the thermal back camera in a smartphone. These studies prove that PPG signals carry the information needed to assess stress levels when data of sufficient quality can be obtained.

Two recent studies have also investigated the ability to detect stress in the field setting based on PPG stress detection models learned in the lab setting [105, 106].

In [105], authors use a combination of several sensing modalities included in the Empatica E2 device for detecting stress in the laboratory and consider context-aware modeling of stress in the natural environment with $(n = 5)$ participants. This work details the accuracy of using PPG and other signal modalities, such as skin temperature and electrodermal attachment from the wrist, as a replacement for more obtrusive chest-based sensors such as ECG or respiration. Our work complements these works by showing that the accuracy of stress inference can be substantially improved by incorporating data quality into the stress inference model.

Our key innovations include the development of a supervised learning-based data-quality indicator for PPG data in the wild and integrating this developed data quality indicator within the stress inference process. Much prior work has also dealt with the problem of corruption of PPG signals owing to hand or wrist motion. They focus on removing PPG segments affected by motion artefacts using conventional or adaptive filtering techniques [107, 108, 109], template matching [110], wavelet transformation [111, 112], independent component analysis [113] and empirical mode decomposition [114]. Much of the existing work on PPG signal restoration is based on motion data collected from finger-based PPG sensors in bedside vital sign monitoring applications where motion is usually limited compared to the natural field environment. We incorporate the established knowledge of motion-induced corruption in PPG signal when estimating heart rate information in the frequency domain using [82]. However, our proposed approach indicates corruption due to motion and other factors in real-life field conditions, such as loose attachment, ambient light, power-line interference, and others. In contrast to most existing work's limited and constrained settings, we collect data in both lab and field settings. The developed signal quality metric allows us to apply relative weighting to different locations of the PPG signal without discarding them altogether. This weighting mechanism diminishes the impact of transient noise and improves the robustness of computed cardiac and heart rate-based features from PPG

signals. As a result, the accuracy and robustness of down-the-line inferences from the computed features improve substantially.

2.6 Smoking Event Detection Using Wrist-only Sensors

A wrist-only smoking event detection model is essential for developing and deploying a smoking lapse risk estimation model from wrist sensors. Existing works on smoking event detection in the field mainly rely on detecting smoking puff events and constructing valid smoking events from the detected smoking puff events. Detection of smoking puffs from wrist-only sensors mostly focuses on hand-mouth of gestures of smoking puffs. Various works have addressed the detection of smoking behavior using wrist-worn inertial sensors. These published studies usually collect lab data from participants with one or more sensors placed into a reference position [57, 58, 59, 60, 61, 55, 56]. However, the utility of smoking detection models developed on lab settings can be limited when deployed to field conditions. Challenges such as variability owing to changes in sensor mounting, sensor placement, and various other factors limit their utility when applied to data collected from the natural environment. Some studies have collected data from the natural field environment [58, 10, 62] with some supervision. However, the developed models make assumptions about the position of wrist sensors on participants' wrists and are vulnerable to the mentioned challenges. Researchers have also explored using other sensors in conjunction with wrist sensors to enhance the performance of smoking puff detection. Respiration sensors (RIP) alone [55, 56] or in combination with inertial accelerometers [10, 115] help smoking puff detection by identifying cigarette smoke inhalation and exhalation characteristics. However, Respiration sensors are chest-worn and place an increased burden on participants to wear a chest-belt device daily. In [61], authors mounted inertial sensing units inside a smart lighter to better distinguish the smoking puffs. Researchers have also advocated using 9-axis IMU units containing quaternions to more accurately estimate the trajectory of hand motion [58]. In [116],

researchers deployed a chest-worn thermal-sensing wearable system that captures spatial, temporal, and thermal information around cigarettes and the wearer to passively detect smoking events throughout the day. Works involving novel sensing schemes to detect smoking behaviors are ongoing to improve any developed models' accuracy and practical utility. Our work focuses on developing a smoking detection methodology using wrist sensors alone. We identify the key limitations of existing smoking detection models when used in the natural field environment and address those challenges to improve the existing works of smoking detection using wrist sensors alone. We develop our methods using smoking data from a natural field environment. Our study setup of training data collection from participants in their natural environment without the supervision of sensor placement and no external control is a first in this research area. We propose an orientation-invariant approach to identify and dynamically align the sensor axes. Finally, we extend the scope of existing smoking puff detection-based models by proposing a novel methodology of smoking episode construction from spurious detected puffs. Our methodology of smoking event detection using machine-learning-based models significantly improves the performance of overall smoking detection. It also adds to our ability to select different operating points based on the application-specific necessity of the developed models.

2.7 Orientation-Invariant Approaches to Dealing with Inertial Sensor Data

Studies employing wrist-worn sensors typically assume sensor attachment at pre-determined positions and orientations with no change over time. This is seldom the case in the natural environment, where many situations induce dynamic changes in sensor placements and orientations. These studies also employ a single inertial sensor (such as Apple Watch) across the whole study, with the developed methods developed and evaluated on data collected from the same sensors. With the ever-increasing popularity of commercial smartwatches fitted with inertial sensors, any developed model must be robust to a change in inertial sensor type and continue operating with the same

level of accuracy. Hence, developing methods invariant to sensor orientation and orientation changes is paramount. Sensor orientation invariant methods have previously been explored in daily activity recognition from wearable sensors [117, 118, 119, 1, 120]. We can broadly categorize the existing works into three separate categories. The first involves collecting representative data from wearable sensors worn in multiple orientations and applying similarity search-based methods to identify the current orientation of the wearable sensor relative to training data [1, 121]. In [1], authors collect labeled training data for 4 possible orientation configurations and do a similarity search based on distribution distance and the distance of the principal component to assign the correct orientation configuration to a window of inertial sensor data. In [121] researchers consider four different orientations for a 3-axis accelerometer on the waist and employ a nearest-neighbor (1-NN) classifier to estimate current orientation. The second category involves the usage of separate sensors along with an accelerometer and gyroscope to orient the sensor coordinate frames to the Earth's coordinates. Researchers used a magnetometer and quaternions to transform sensor coordinates to Earth's coordinates [119, 122]. These two categories of work do not fall within our scope since they require additional sensing schemes [119] and labeled training data [1] to achieve orientation invariance. The final category of work related to orientation-invariant inertial sensor data processing involves transforming the inertial sensor data streams into an orientation-invariant representation. A straightforward method of obtaining orientation invariance is calculating each tri-axial sensor's Euclidean norm (magnitude) and using the magnitude sequence instead of the individual axis components [123, 124, 125, 126, 127, 128]. The magnitude remains the same even when the sensor is placed at different orientations. We follow this methodology for activity detection from wrist-worn accelerometers in Chapter 4. However, magnitude time series do not preserve the fine-grained information of the individual sensor axes necessary for detecting complex activities such as smoking. In [129, 130], authors

estimate the gravity vector's direction by averaging the acceleration vectors in the long term. Next, the amplitude of the acceleration along and perpendicular to the gravity vector are used for activity recognition. This method is analogous to transforming the tri-axial sensor into a bi-axial one. In [117], authors propose transforming 3-axis sensor data to a 9-axis orientation-invariant time-domain sequence. Existing works that propose transforming the inertial sensor data into an intermediate representation have also employed Principal Components Analysis [131] and Singular Value Decomposition [117] to transform sensor data into the same number of dimensions as the original data. Using the transformation-based methods for step-by-step explainable smoking detection using wrist-worn inertial sensor data has some limitations. The individual sensor axes lose meaning when transformed into a different system of coordinates. For example, in smoking detection, the candidate puff segments are generated from changes in the gravity axis, and spurious segments are filtered out based on roll values which indicate the angular rotation around the gravity axis [10]. We aim to preserve and utilize this domain information related to the smoking detection problem in proposing an orientation-invariant method of smoking detection. The existing methods also assume that the sensor remains the same across the duration of the study. The transformations assume that the x,y, and z axes will remain the same for a single sensor type. However, variability in sensor configurations due to changing sensor hardware or firmware can switch the individual axes. In contrast to the related works, our orientation in-variance approach tackles two problems. We propose methods to identify the configurations of a given wrist-worn sensor using the distribution of the accelerometer sensor signals during moments of walking. Next, we propose to align the sensor axis corresponding to the gravity line in real time. We base our methods on addressing the limitations facing existing smoking detection methods using wrist-worn inertial sensors in the natural field environment.

2.8 Learning from Sparse Positive-only Labels

Traditional supervised classifiers usually need concrete positive and negative samples for training. In our case, we only have no labels for the negative class (low-risk state) and only a few positively labeled samples from smoking detection and confirmation through self-reports. For such scenarios, a different learning framework called *Positive-Unlabeled (PU)* have been developed [132, 133].

In the classical PU learning algorithm [134], a standard binary classifier is trained from the nontraditional positive-unlabeled setup. They show that a classifier trained on positive unlabeled examples learns probabilities that differ from the actual conditional probabilities of being positive by only a constant factor, equivalent to the likelihood that a positive sample is labeled in the given data set. Using different weights for false Negatives vs. false Positives in training has also been proposed for solving the classical PU-learning problem. For instance, the biased SVM approach in [135] solves the PU-learning problem by using soft margin SVM while giving high weights to false negative errors and low weights to false positive errors. The authors also used weighted logistic regression models to classify text by considering all unlabeled instances as members of the negative class with appropriate weights [136]. However, these classic PU learning algorithms work only under the strong assumption that the set of labeled examples is a uniformly random subset of the positive examples (or the positive-label samples are 'selected completely at random' (SCAR)). For scenarios like ours where the SCAR assumption does not hold, the PU-bagging [63] or ensemble PU learning [137] have been proposed. The idea is to estimate a series of classifiers on datasets obtained by perturbing the original training set through bootstrap re-sampling with replacement. Finally, an aggregation technique is applied to combine these classifiers. We adopt PU-Bagging [63] to train our smoking risk estimation models similar to [24]. We augment the training of base classifiers at each iteration of bagging by using a proposed novel loss function. Our proposed loss function is related to automatically estimating the

proportion of positive data instances within the unlabeled set of instances. Often known as the class prior, this proportion is assumed to be known for training models in positive-unlabeled learning [138]. Works relating to estimating this proportion falls within the Mixture Propagation Estimators (MPEs) category [139, 140, 141]. MPEs estimate the fraction of positives among the unlabeled examples, and PU-learning incorporates this estimate into a scheme for learning a binary classifier [141]. Post-processing approaches of MPE and PU learning employed in DedPul [140] depend on heuristics based finetuning of the output of positive-unlabeled classifiers. Authors also proposed Best Bin Estimation (BBE) first to produce a consistent estimate of the mixture and then integrated it with an iterative model training approach of PU Learning [141]. The use of non-convex loss functions [142] and added regularization [143] have also been employed to learn classifiers for PU learning. We complement these works by proposing a metric learning-based loss function to enable learning accurate PU classifiers using the PU-Bagging [63] approach of model learning. Our loss formulation includes the mixture proportion term as a hyperparameter and provides pathways for selecting the optimal value in training times.

2.9 Chapter Summary

This chapter provided a detailed overview of the literature concerning our problem. We first categorized our overall goals to construct multiple major sub-problems. We described the state-of-the-art methods or strategies for each problem. From now on, we present our completed works in Chapter 3, Chapter 4, Chapter 5, and Chapter 6. The methodologies implemented in these following chapters are inspired by the latest works mentioned herein.

Chapter 3

mRisk: Continuous Risk Estimation for Smoking Lapse from Noisy Sensor Data with Incomplete and Positive-Only Labels

3.1 Introduction

Interventions delivered on a mobile device are an important tool to improve health and wellness via behavior change such as for smoking cessation. Decades of research in pharmacological and behavioral intervention methods have improved the success rate of quit attempts, but they still hover near 30% [144]. Knowing when the participant is at-risk of an adverse behavior can enable the exploration of whether and how well delivering targeted interventions at moments of risk can improve efficacy. For example, [145] presented a context-aware method to deliver timely interventions by sensing the exposure to geolocation-based smoking cues.

To detect the *high-risk* moments of an imminent adverse event, it is important to identify the dynamic risk factors that influence the occurrence of the adverse event. Prior research [25, 23, 146] has shown that these risk factors can be divided into two categories. First are the 'external' stimuli, i.e., environmental/social cues conducive to lapse (e.g., proximity to a bar or seeing others smoke may increase the risk of a smoking lapse). Second are the 'internal' stimuli such as stress or craving that may increase an individual's vulnerability to lapse. Depletion of coping capacity during exposure to risk factors may result in a lapse.

Behavioral science suggests that just-in-time interventions, aiming to prevent a lapse, should adapt to both dynamically varying internal and external factors to provide optimal support at the right moment [146]. The emergence of sensors in wearables and smartphones has made it possible to passively detect dynamic changes in internal risk factors (e.g., stress [15, 147] and craving [30, 70]). Dynamic changes in the external risk factors for smoking lapse can also be detected passively using GPS and activity sensors (e.g., visits to smoking spots [32]). Deriving a composite risk score that reflects the

dynamically varying levels of risk continuously can provide new opportunities to optimize both the timing and contents of interventions via micro-randomized trials [148].

Substantial work has been done in estimating risk scores for other kinds of adverse events. They include mortality [35, 36, 37, 38], ICU admission [33, 34], disease onset [39, 40, 41, 42, 43], fire hazard [44, 45, 46], flood [50], wildfire [51, 52], and road accidents [47, 48, 49]. The use of deep learning models helps obtain a composite risk score that encodes the underlying collective predictive power of all the input risk factors. For training and testing these models, carefully curated and labeled input data with timestamps of adverse event occurrences are used. All data not labeled to correspond to an adverse event are usually treated as negatively-labeled (i.e., low-risk). For example, when predicting mortality in ICU from large-scale electronic health records data (e.g., MIMIC-II), each of the 4,000 patients is either in the mortality (534 in Class 1) or the survival class (3,466 in Class 2) [35].

Estimating a composite risk score for adverse health-related behaviors poses three new challenges. First, continuous sensor data collected from wearables and smartphones to capture risk factors of adverse behaviors in the natural environment are usually noisy and incomplete [149]. This may be due to lack of firm attachment (e.g., proximity of pulse plethysmography (PPG) sensor to the skin in smartwatches that are used to detect stress and craving), intermittent noises (e.g., motion-induced deterioration of PPG data due to frequent wrist movements), and confounds (e.g., elevated physiology during recovery from physical activity may be confused with stress response). Second, for adverse behavioral events such as a smoking lapse, capturing the precise timing of each smoking lapse may not be feasible, as sensors may not be worn at the time of a lapse or the lapse events may not be accurately detected due to the imperfection of machine learning models that are used to detect smoking events via hand-to-mouth gestures [10]. Therefore, only a few positive events (i.e., smoking lapse in a cessation attempt) are available. Third, confirmed negative labels can be assigned to a block of sensor data

corresponding to a prediction window only if the entire time period is confirmed to have no high-risk moment. As not all high-risk moments may result in a lapse, labeling a block of sensor data to the negative class is difficult for such events.

In this chapter, we address each of the three challenges noted above. We first encode the noisy sensor data in the form of events that represent the psychological (e.g., stress), behavioral (e.g., activity), and environmental contexts (e.g., proximity to a smoking spot). Second, each of these contexts has substantial diversity in their representation (e.g., frequency, duration, type, etc.). We compute their homogeneous statistical representations to use them in training deep learning models. Third, we explore two approaches to succinctly capture the historical influence of recent and past events (i.e., substantial change in any context) to make deep learning models efficient. In the first approach called *Deep Model with Recent Event Summarization (DRES)*, we summarize the influence of recent and past events via features. In the second approach called *Deep Model with Decaying Historical Influence (DDHI)*, we explicitly encode the influence of recent and past events as an exponentially decaying function over time. We refer to both models as *mRisk* model choices. Fourth, we address the challenge of sparse and positive-only labels via the *Positive-Unlabeled (PU)* framework, which allows for model training with positive-only labels. However, *PU* frameworks usually train models by giving higher weights to the positive samples and use a spy dataset (that has a small number of both positive and negative samples) for evaluation [150]. But, we do not have access to even such a small spy dataset. Therefore, we design a new loss function (called *Rare Positive (RP)*) to train the *mRisk* model choices and use the concept of the rarity of the positive class for evaluation.

We train and test the two models on a real-life smoking cessation dataset. We evaluate the performance of the two models via the risk characteristics they produce and their ability to create intervention opportunities prior to each confirmed smoking lapse moment. We find that 85% of lapses can be intervened upon with about 5.5

interventions per day. By analyzing the risk dynamics around lapse moments, we discover that risk usually peaks 44 minutes prior to a lapse. Finally, we use *SHAP* [151] to explain the influence of different contexts on lapse risk and find that recent visit to a smoking spot has the highest influence on risk, followed by stress.

3.2 Smoking Cessation Study and Data Description

We introduce smoking cessation, describe the smoking cessation study, and the resulting data used in modeling. The Institutional Review Board (IRB) approved the study, and all the participants provided written consent.

3.2.1 Smoking Cessation Research

Smoking is the leading preventable cause of mortality, causing 7 million deaths globally each year [152]. Therefore, extensive research has been done to support smoking cessation and to understand the smoking lapse process to improve rates of successful quitting. When a smoker attempts to quit smoking (i.e., abstain), withdrawal symptoms due to nicotine deprivation trigger several physiological and behavioral changes such as increase in stress, anxiety, concentration impairment, and craving [23, 67]. These changes can be further accentuated by certain situational or environmental influences such as exposure to smoking cues (e.g., proximity to a cigarette point of sale) or social triggers (e.g., drinks with friends) [23, 153]. These physiological and/or situational events constitute a *high-risk* situation for a smoking lapse. Individuals who are unable to cope with the acute challenges of *high-risk* situations, transition from abstinence to a smoking lapse [154]. In most cases, the first lapse eventually leads to full relapse [155, 22]. To capture risk factors for a smoking lapse that can be passively detected from wearable sensors and used for continuously estimating lapse risk, we conducted a new smoking cessation study.

3.2.2 Participants

Participants were recruited in a number of ways. First, recruitment flyers were posted in public areas such as college campuses, community clinics, churches, and in local restaurants and bars in Houston. Advertisements were placed in local newspapers and on radio. In person recruitment was implemented as needed to promote enrollment, or if requested by groups or institutions that have a population who is likely eligible and interested. The recruited participants went through the informed consent process during their initial (baseline) lab visit.

We use data from 170 enrolled participants (76 female), all 18+ years of age, with a mean age of 49.158 ± 12.99 years. All participants were African-American, residents of a city in the USA, smoked at least 3 cigarettes per day, and were motivated to quit smoking within the next 30 days of the start of the study. All of them agreed to wear the sensor suite. Participants were excluded if they had a contraindication for the nicotine patch (e.g., participants at risk of heart attack, angina, and other related health problems), active substance abuse or dependence issues, physically unable to wear equipment, pregnant or lactating, or currently using tobacco cessation medications.

3.2.3 Study Protocol

Interested participants were invited to an in-person information session where they were provided with detailed information about the study. Once enrolled at the baseline visit, participants picked a smoking quit date. They visited the lab during which they were trained in the proper use of the sensor devices and how to respond to questionnaires in the form of Ecological Momentary Assessments (EMA) via a study-provided smartphone. They wore the sensors for 4 days during the *pre-quit* phase.

On their set quit date, participants returned to the lab. Then they wore the sensors for 10 more days during the *post-quit* (or *smoking cessation*) phase. At the end of 10 days (14 days from the study start), participants returned to the lab and underwent biochemical verification of their smoking status. The participants were

compensated for completing in person visits — $30 each for Visits 1, 2, and 3, $80 for Visit 4, and $60 for Visit 5. They were further compensated at the rate of $1.25 for completing each smartphone survey if they wore the on-body sensors and/or collected usable sensor data at least 60% of the time since the last phone survey, and $0.50, otherwise for completing each smart phone survey. The participants were also reimbursed for parking or bus tokens to defray the cost of traveling to the project site.

3.2.4 Wearable Sensors and Smartphone

Participants wore a chest band (AutoSense [72]) consisting of electrocardiogram (ECG) and Respiratory Inductive Plethysmography (for respiration) in their natural environment for up to 16 hours per day. We use the physiological data for continuous stress inference. To capture physical activity context, AutoSense included a 3-axis accelerometer. The participants also wore a wristband with 3-axis accelerometers and 3-axis gyroscopes on both wrists. Participants carried the study-provided smartphone with the open-source mCerebrum software [156] installed. The study smartphone was used to communicate with the wearables and collect self-reports via EMAs. The smartphone collected GPS data continuously at a rate of 1 Hz. We use the GPS data for detecting significant locations. The GPS data was extracted from the phone at the end of the study. All data from wearable sensors, EMAs, and GPS were stored in a secure server with the open-source Cerebral-Cortex [157] software installed.

3.2.5 Determining the Smoking Lapse Time

The participants reported smoking events via Ecological Momentary Assessments (EMA). For uniform coverage, the day was divided into 4 blocks. The first three blocks consisted of 4 hours each, with remaining time assigned to the last block. In each block, up to 3 EMAs were triggered with a minimum separation of 30 minutes between successive prompts. Irrespective of the source (random or triggered by the detection of stress or smoking), each EMA included the following questions, *'Since the last assessment, have you smoked any cigarettes?'*, *'How many cigarettes did you smoke?'*,

'How long ago did you smoke the cigarette?', and *'How long ago did you smoke the most recent cigarette?'* and *'How long ago did you smoke the first cigarette?',* if multiple cigarettes were smoked.

The precise time of smoking lapse is needed to label the corresponding sensor data to belong to a positive class. To pinpoint the time of a smoking lapse, we utilize the puffMarker [10] model that detects smoking episodes using a machine learning model trained to identify deep inhalation and exhalation from a RIP (Respiratory Inductive Plethysmography) sensor and hand-to-mouth gestures from 6-axis inertial sensors (3-axis accelerometers and 3-axis gyroscopes) worn on both wrists. But, some smoking episodes may not be detected (due to model imperfections, sensor non-wear, etc.) as well as some non-smoking events (e.g., eating popcorn that involves similar hand-to-mouth gestures) may be falsely detected as smoking episodes. Hence, we also use smoking labels provided by the participants in EMA's. For training the *mRisk* model, we only use those detected smoking episodes that are also supported by participants' self-reports in EMAs.

The time point from which a smoker is actively attempting to abstain from smoking is called the *quit time*. Although any smoking event after quitting is considered a *smoking lapse*, situations when a newly abstinent smoker promptly resumes abstinence after the initial smoking event are regarded as slip-ups. The resumption of usual smoking after quitting is considered a full relapse, and end of the current quit attempt. The time interval between quitting and the onset of full relapse is the *abstinence period*. Based on prior research [158], we consider three (3) consecutive days of smoking after the first smoking lapse as the onset of full relapse, and end of the abstinence period. We use all confirmed smoking events during the abstinence period as the positive class.

3.2.6 Data Selected for Modeling

Some of the physiological data was not of acceptable quality due to sensor detachment, loose attachment, persistent and momentary wireless loss between the phone and the sensor. Using the methods presented in [149], we identify sensor data of acceptable quality and use them in our modeling.

Out of 170, eight (8) participants completed the pre-quit phase, but did not return for the post-quit. Additionally, eleven (11) participants were unable to complete the entire study. Hence, we were left with 151 participants who completed the study. As we use cross-subject validation, we ensure uniformity and sufficiency of continuous inference data. Therefore, we select participants based on the following two criteria. First, the participants have a minimum of three hours of stress and activity inferences each day (this produces sufficient stress and activity data for model development). Second, the participants have GPS data for consecutive days across the pre-quit and post-quit days (this allows us to derive sufficient location history for model development). As a result, 59 participants were excluded. The 92 remaining participants wore the AutoSense chest band for an average of 14.63 hours per day. From these participants ($1,012$ person-days), we obtain a total of $11,268$ hours of stress data (11.13 hours each day) and $14,066$ hours of activity data (13.89 hours each day) for model development. We also obtain a total of $17,569$ hours of location data (17.36 hours each day) and $3,719$ completed EMAs (out of $5,210$, 71.38% completion rate). Out of 92 selected participants, 56 have puffMarker-detected lapses also confirmed by EMA.

3.3 Problem Setup and Formulation

3.3.1 Problem Formulation

Our goal is to develop a model that can process the continuous data from sensors in wearable devices and smartphones and obtain a score that can indicate the risk of lapse at each moment, providing new intervention opportunities to maintain smoking abstinence. To formulate our problem, we introduce some terms and definitions.

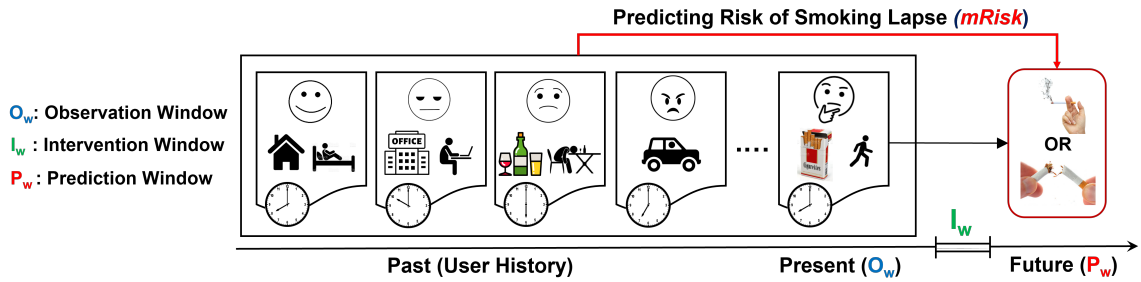


Fig. 3.1: Internal state and external cues from an observation window and prior to it are used to estimate the risk of a smoking lapse during the prediction window. The intervention window between the observation and prediction windows gives an opportunity to deliver an intervention.

Following the setup from [159], an *Observation Window* (O_w) is a fixed-length time interval such that data collected in this time window and any historical context prior to it are used to estimate the likelihood of the target adverse event occurring in an upcoming *Prediction Window* (P_w) (see Figure 3.1). We introduce a gap after the end of an observation window and before the start of a prediction window, which we call the *Intervention Window* (I_w), where an intervention might be beneficial in preventing the adverse event predicted to occur in the *Prediction Window*. We slide all windows over the continuous stream of sensor data with an offset of 1 minute.

Problem: Given the time series of sensor data and the timing of some smoking lapses from a population of abstinent smokers, train a model $\mathcal{M}$ that can accurately estimate the risk of lapse in a prediction window P_w for an abstinent smoker, using the sensor data observed up to and including the corresponding observation window O_w, such that the proportion of all prediction windows estimated to have a *high-risk* of lapse is minimized.

3.4 Robust Computation of Psychological, Behavioral & Environmental Context

We apply existing trained models to accelerometry, ECG, Respiration, and GPS data to capture the following psychological, behavioral, and environmental contexts of users, as *continuous inference streams* (see Figure 3.2a).

Stress: As stress can influence a smoking lapse, we obtain a continuous assessment of physiological stress arousal by applying the cStress model [15]. cStress computes a set of features from one-minute windows of ECG and respiration data and produces a likelihood that the user is exhibiting stress arousal in the captured physiological response. We apply the cStress model on our smoking cessation field study ECG and respiration data to generate stress likelihood every five seconds from overlapping, i.e., sliding minute windows to get a smoother time series. The cStress model produces a value between 0 and 1 that we call our *stress inference stream*.

We handle short episodes of missing data in the stress inference stream (due to noisy data, confounding physical activity, or recovery from physical activity), by applying the k-nearest neighbor-based imputation [160].

Activity: Movement such as transition from inside to outside of a building can expose a user to potential environmental triggers of a smoking lapse (e.g., corner of a building designated as a smoking spot). Therefore, we obtain an assessment of non-stationary or *active* state for each minute. We utilize the 3-axis accelerometer sensor embedded in AutoSense for activity detection (of the torso) using the model presented in [149].

Location History: Change in location can expose a user to major environmental cues such as tobacco point of sale or bars. Therefore, we obtain a continuous assessment of change in a participant's location. We adapt the context mining approaches used in [32] to derive location history, dwell places, and transitions from raw GPS data. First, we de-noise the GPS data via median filtering [161] as the gap between

consecutive GPS points is much less than fifty meters even at a speed of 100 kilometers per hour due to the sampling rate of 1 Hz in our GPS data. We perform median filtering by substituting a GPS sample point with the median of temporal predecessor points from a window length of 2 minutes (i.e., 120 predecessor points). This step produces a continuous inference stream of location history (time, latitude, and longitude). Finally, we employ spatio-temporal clustering to derive the start and end times at dwell places (both significant and transient) or transition from one place to another.

3.4.1 Robust Representation of the Current Context

The current context, i.e., measures of stress, activity, and location history inferred from the observation window, are heterogeneous as they are sampled at different rates, and transitions can happen dynamically. Although not as noisy as the raw sensor data they are derived from, they still suffer from noise, discontinuity, and rapid variability due to model imperfections, sensor non-wear, data quality issues, and confounding events.

To address these issues and obtain a homogeneous and robust representation of the current context that can be used to train a deep learning model, we compute statistical features of continuous inference streams. Such aggregate statistical measures have more robustness to noise compared to raw inferences themselves.

We use 13 statistical functions to compute features from the stress stream. These functions compute the elevation (80^{th}, 90^{th}, and 95^{th} percentiles), reduction (20^{th}, 10^{th}, and 5^{th} percentiles), dispersion (interquartile_range and skewness), central tendency (median), shrinkage (range between [20^{th}, 10^{th}] and [20^{th}, 5^{th}] percentiles), or accumulation (range between [80^{th}, 90^{th}] and [80^{th}, 95^{th}] percentiles) from a window of inferences. Given an observation window (t_i, t_{i+w}) of length $w = |O_w|$ minutes, we have a maximum of $12 * w$ stress state data points, since an assessment is produced every 5 seconds. We compute stress features as follows. Thirteen (13) statistical features are obtained from the stress stream from t_i to t_{i+w}. The same functions are also applied to the consecutive difference between the successive stress likelihoods in the window. To

account for day-specific within-person variability, we compute the statistical features (called *baseline features*) from day-long stress stream up to t_{i+w} (we use *until_obs* to abbreviate 'until the end of observation window'). Finally, we capture the average deviation of stress (from the daily mean) at the current window.

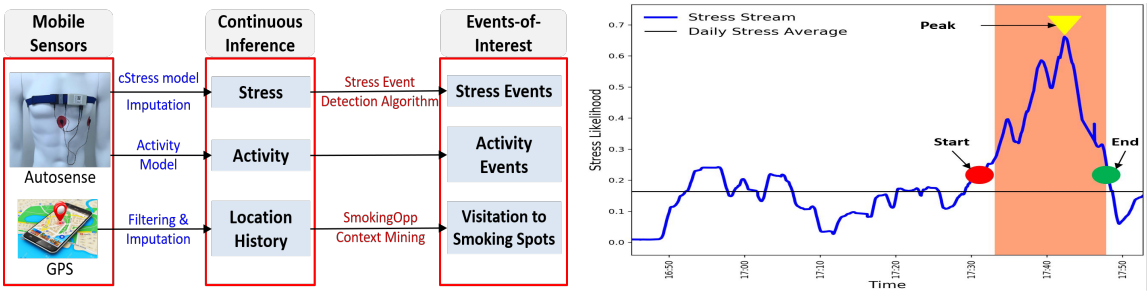
We compute the fraction of time active in the current window from the activity stream in an observation window. From the location stream in the observation window, we compute a binary indicator to check if the current location is a smoking spot (=1) or not (=0). Next, we compute the distance to the nearest smoking spot. Finally, we compute the fraction of time spent stationary at a place, the fraction of time spent in transition in the current window and the current speed.

In total, we compute *46* features from the three inference streams, called the *Continuous Inference Features*.

3.4.2 Encapsulating History via Events-of-Influence

A key question for the *mRisk* model is how to describe the influence of context on lapse risk over time. Continuous measures of factors such as stress are likely to have only proximal impact on risk, which is modeled by the temporal interval between the observation and prediction windows. However, significant contextual events, such as period of extremely high stress, may have a degree of influence over a significantly longer interval of time. We, therefore, define events of influence, which are specific contextual events occurring at discrete moments in time, and model their influence on risk prediction. Hence, our next challenge is how to succinctly capture the influence of these historical contexts so that the model may be able to estimate the degree of their influence and how it may wane over time. We encapsulate the historical contexts by computing *events-of-influence* streams (see Figure 3.2a) from the corresponding continuous inference streams as was done in [24].

Event-of-influence stream is a sequence of irregularly spaced events derived from the continuous inference stream. An event represents a location in time, which likely



(a) Sensors and extracted events used for model development.

(b) A stress stream with a stress event.

Fig. 3.2: (a) Sensors and extracted events used for model development (b) A stress stream with a stress event

impacts the participant's current and future actions. Each event comprises of one or more attribute values, a start time, and an end time, represented as *<list of values, start, end>*. The type of attribute values in different events-of-influence stream can be numerical, binary, or categorical. Specifically, we compute three events-of-influence streams.

Stress Events

The model presented in [160] applies a moving average convergence divergence (MACD) method to detect the increasing or decreasing trend and the inflection point (or the peak) in the stress likelihood time series based on short-term and long-term exponential moving average. This method clearly marks each stress event's *start* and *end* times, defined as the interval containing the increasing-trend interval followed by a decreasing-trend interval. Each stress event has the following attributes — the *stress duration*, which is defined as the time interval between the *start* and *end* of a stress event (in Figure 3.2b, we observe a stress event of 14.75 minutes) and the stress density, which is defined as the area under the stress stream divided by the stress duration (in Figure 3.2b, we observe a stress event with density of 0.445). Each stress event is represented by *<stress density, stress duration, start, end>*. Finally, the model applies a threshold based on the stress density to determine which events are stressful and which are not. We note that stress or non-stress events are only detected from those segments

of sensor data that are of acceptable quality and not confounded. In our data, on average, we detect 2 to 9 stressful events per day with a mean density of 0.242 and a mean duration of 10.747 minutes.

Activity Events

We employ the following approach to detect the activity events from the activity stream. We cluster the contiguous active and stationary windows together to construct the active and stationary events, respectively. Each activity event is represented as *<binary indicator of 1, duration, start, end>*. In our data, on average, we detect 12 activity events per day with a mean duration of 2.70 minutes.

Visitations to Smoking Spots

Smoking spots are those places where participants are observed to have smoked, smoking is usually allowed, and cigarettes are available. We employ the spatio-temporal context mining methods described in [32] to locate the two categories of smoking spots (personal and general smoking spots) from participant's location history and smoking patterns.

Visitations to smoking spots are recorded as events-of-influence. We adapt the method from [32] to detect a visitation to a smoking spot (when a participant dwells for at least 6.565 minutes with the distance of $30m$ from the centroid of a smoking spot). Each visitation to smoking spot event is represented as *<semantic type, duration of stay, start, end>*, where we consider the following *semantic types* for our analysis, smoking outlet, retail store, gas station, or a bar (usually cigarettes are available at these location types), *start* is the arrival time to and *end* is the departure time from the smoking spot. Duration of stay at a smoking spot is computed as the difference between the departure and the arrival time. In our data, on average, we detect about 1 visitation to smoking spots per day with a mean stay duration of 12.48 minutes.

3.5 *mRisk*: Modeling Imminent risk of lapse

In developing the *mRisk* model, we aim to discover a suitable representation of the event-of-influence time-series and find the role of continuous context variables within the observation window in predicting the lapse risk. We first opt for traditional feature representation of the event time-series. We use several features from [24] to summarize the influence of events on modeling the lapse risk phenomenon. We term this model *Deep Model with Recent Event Summarization (DRES)*. In an alternate approach, we hypothesize that events have a decaying influence over time on the risk of lapse. We explicitly model the decaying influence using exponential decay functions. Furthermore, we incorporate knowledge from the patient sub-typing domain [162] to enable end-to-end model learning, with both dynamically changing instance variables and static variables reflecting an aggregate phenomenon. We refer to this model as *Deep Model with Decaying Historical Influence (DDHI)*.

3.5.1 Deep Model with Recent Event Summarization (DRES)

For the *DRES* model, we represent the *event-of-influence* Using features. These features complement the statistical features obtained from the observation window described in Section 3.4.1. The architecture of the *DRES* model also includes the encoding of the recent past with a stacked observation window-based design. We present the features used to summarize the events-of-influence and the model architecture in the following.

Events-of-Influence Representation using Features

We use *15* events-of-influence features to capture the temporal dynamics of the psychological, behavioral, and environmental events from the recent past. These features are extracted from three events-of-influence streams corresponding to an observation window.

- **Stress Events:** We compute *average duration & density of stress events within current window, time since the previous stress event, duration & density of the*

previous stress event. Additionally, we compute the *average duration & density of stress events,* and *fraction of time in the stressed state* until the observation window.

- **Activity Events:** We compute *time since the previous activity event* and *duration of the previous activity event.* Additionally, we compute the *average duration of activity events* and *fraction of time in an active state* until the observation window.
- **Visits to Smoking Spot Events:** We compute *time since last visit to a smoking spot* and *average duration of stay at smoking spots.* We also compute the *fraction of time spent at smoking spots* until the observation window.

Feature Set

We compute 61 total features from the continuous inference and event-of-influence streams. We also include the *hour of day* (using one-hot encoding) as a feature based on prior work [68] which shows time may affect the occurrence of a smoking lapse. In total, we compute 62 features per observation window for the *DRES* model development. We adopt per-participant standardization to account for between-person differences and introduce feature baselines to incorporate within-person variability or individual biases in features.

DRES Model Architecture

The idea behind the *DRES* model is that all the features computed in each observation window can be represented in a time-lagged fashion to accurately estimate the risk of lapse likelihood in the prediction window. Figure 3.3 shows the overall architecture of *DRES* model. Here, X_t refers to the feature vector computed from an observation (i.e., time-lagged) window starting at time t. We use the tabular features $n_f = 62$ from each observation window. Next, we stack features from n_l previous observation windows, with the size of the input instance being $n_l \times n_f$. The n_l observations provide information on the temporal evolution of features in the recent past

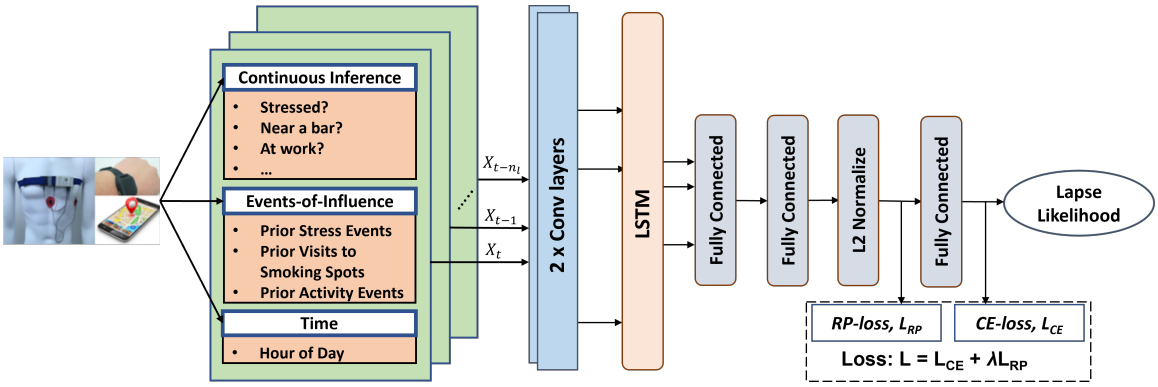


Fig. 3.3: Overall architecture of the Deep Model with Recent Event Summarization (DRES)

(hence, the term *Recent* in *DRES*). The efficacy of *DRES* model depends on the ability of hand-crafted features to properly encapsulate the spatial-temporal-behavioral cues useful in predicting lapse. Since *DRES* model utilizes regularly sampled feature vectors stacked together in time, we use a simple Convolution plus LSTM architecture. The model's overall architecture consists of two convolutional layers, one recurrent LSTM layer, three fully connected layers, and a single node sigmoid layer. The convolution layers help to extract micro-features in a local neighborhood followed by an LSTM layer which captures temporal patterns of the micro-feature sequence. The recurrence in the LSTM is operating along the n_l lagged windows. The penultimate fully connected layer is followed by an L_2 normalization layer to normalize the input vectors to unit norm. Finally, the output of the final fully connected layer is passed through a single node with a sigmoid activation function to generate the lapse likelihood.

3.5.2 Deep Model with Decaying Historical Influence (DDHI)

For the *DDHI* model, we explicitly model the decaying influence of a past event. For the current context, we continue to use the statistical features from Section 3.4.1. But, we observe that the proposed event of influence features in the *DRES* model rely heavily on the usefulness of specific features calculated and are limited to only incorporating the most recent past events and the average information. We propose an alternative event encoding approach that allows for encoding of multiple past events and

enables the model to learn from not only recent events but also the accumulated effect of past on participants' psychological and contextual state without explicit feature engineering. First, we provide the rationale for the development of our proposed methodology. Next, we formally define the encoding procedure and the various design choices involving the model architecture.

Modeling Rationale

Lapse risk may be influenced by not only recent internal and external events but also by the accumulated history of exposures, with the influence waning over time. To model this behavioral element, we need to efficiently represent the stimuli received by the participants from earlier time points. In estimating the risk of imminent adversarial behavior, our goal is to directly account for the current influence of past events, weighted by their position in time. The event-of-influence streams are also unique in their discrete nature of non-aligned multi-modal observation. The unique aspects of event modeling make it challenging to directly apply the current deep-learning modeling approaches to our scenario.

Modeling with time-series data requires encoding previous states as time progresses. Long-Short Term Memory Networks (LSTMs), Time-Aware LSTM networks [162], and attention-based LSTMs [43] have all been used successfully to model time series data. They have produced state-of-the-art results in time-series problems such as mood forecasting [163], mortality prediction [164], and intervention delivery [159]. *Transformers* [165] with the self-attention mechanism has proven highly successful in modeling long-term dependencies for sequential data, enabling learning of large sequence models for multivariate long-term forecasting [166].

In our case, to capture the historical influence, the model needs to learn from the events-of-influence streams. Different events-of-influence streams have observations at different times with scant alignment between them. To properly capture the historical influence, we need to be able to learn from these multiple irregular time series from

further in the past. We also need the model to learn from mutual interaction of multiple past event types by aligning their decaying effect in a future time, which is not yet handled well in current models. To efficiently model long-range temporal interactions of irregularly sampled non-aligned observations, we want a model where the temporal delay can be explicitly designed because it's a key aspect of our problem.

Therefore, we propose a decay-aware temporal embedding of heterogeneous past events to encode their residual effects in predicting the lapse risk. We represent each event using a standard set of attributes and use the encoding approach to propagate the effects of past events. In this way, we aim to create a temporal projection of any past event in times of future inference. Our proposed methodology transforms event data using an exponential decay function before feeding it to an LSTM layer. The LSTM layer provides a simple way of handling the time-dependency within the current observation of limited length. To estimate effective exponential decay factors and weights for different event attributes, we adopt the patient phenotyping approach from the EHR domain [162]. We analyze the feasibility of grouping our participants using global aggregate context variables from the pre-quit period and use the grouped representation as a key input variable in the model.

Decay-aware Temporal Encoding of Heterogeneous Events

We represent a single event using a vector of k attributes, $B = [\beta_1, \beta_2, ..., \beta_k]$ alongside the time of event t. For example, stress events can be represented using, the time of event t, density β_1, duration β_2, peak amplitude β_3 and other factors. These attributes are determined by the event type. For example visit to smoking spots is an indicator event with no density information present. We represent a single event of type e (e.g., stress, visit to smoking spots, activity) using the tuple (t, $B^e = [\beta_1^e, \beta_2^e, ..., \beta_k^e]$). We aggregate the contributions of k different attributes of an event in a single numerical value using a linear function,

$$f(B^e) = \frac{1}{k} \sum_{i=1}^{k} \mu_i^e \beta_i^e \tag{3.1}$$

Here, μ_i^e is the weight coefficient associated with the i^{th} attribute, β_i^e . We standardize each attribute to be within the range $[0, 1]$ and estimate the weight coefficients using sigmoid function — $0 \leq \mu_i^e \leq 1$. The division by the number of attributes k ensure that $0 \leq f(B^e) \leq 1$ for all event types with different number of attributes.

To represent an event from the past (t_1, B^e) at a future time $t \geq t_1$, we assume an exponential decay function of a constant rate α^e with $f(B^e)$ representing the initial quantity from (3.1). Thus, the contribution of the event from time t_1 at a future time t becomes $f(B^e)e^{-\alpha^e(t-t_1)}$. Exponential models are widely used to model decay in natural phenomenon such as drug absorption [167], recovery times from physical activity [160], among others.

Thus, given n past events of Type e, $(t_1, B_1^e), (t_2, B_2^e), (t_3, B_3^e), ..., (t_n, B_n^e)$, we aggregate the effects of all past events at time t as $\hat{s}_t^e$ with

$$\hat{s}_t^e = \sum_{k=1}^{n} f(B_k^e)e^{-\alpha^e(t-t_k)}I(t \geq t_k), \tag{3.2}$$

where $I(t \geq t_k)$ is an indicator function equal to 1 if $t \geq t_k$ and 0, otherwise. The parameter α^e controls the rate of decay of an event of type e as we progress in time. We directly feed the time-series $\hat{S}_{t-w:t}^e$ of different event types (stress, activity, smoking spot visits) to the model together with statistical features from the current window O_w to allow the model to learn from accumulation of past events. Since Equation 3.2 can be computed at any time in future, we can maintain the regular time intervals required for a simple LSTM to operate on.

Our embedding depends on effective estimation of the parameters α, $[\mu_1, \mu_1, ..., \mu_k]$ for each event type. We assume that these three parameters act as variables specific to global contexts. For example, we assume that the decaying rate of influence of stress on lapse risk does not change from one stress event to another and is similar for a set of homogeneous participants (i.e., phenotype). Thus, estimation of the

parameters α, $[\mu_1, \mu_1, ..., \mu_k]$ depends on identifying the degrees of freedom on which each participant is homogeneous.

Phenotyping Participants for Parameter Estimation

Patient sub-typing is grouping of patients to address the heterogeneity in the patients, to enable precision medicine where patients are provided with treatments tailored to their broadly unique characteristics [162]. We group participants based on observations from the pre-quit period so that the model can be applied to a user right from the moment they quit when no post-quit data is available. The features we use include gender, age, average stress density, duration and count per day before quitting, average frequency and duration of visits to smoking spots prior to the quit period, and average activity event count and duration per day. We term them phenotype features since they provide relatively stable information (i.e., trait) about the participants. We aim to cluster the participants into a small number of groups.

Clustering: Our clustering of participants based on their phenotype features is guided by three questions — *which clustering algorithm to use, which features contribute most towards a grouping of the participants,* and *how many clusters are appropriate.* We experiment with partition-based traditional k-means algorithm and hierarchical clustering approaches. Both methods perform similarly in our data. We vary the number of clusters for obtaining the most appropriate clustering. For identifying the features which are most useful in grouping the participants into different clusters, we select silhouette score [168] as the evaluation criterion. First, we re-scale the features to fall within the same range between 0 and 1. Next, we measure the silhouette score of removing a single feature at every iteration and remove the feature which contributes negatively toward the overall clustering. This recursive feature elimination process allows for identification of the most important features necessary for grouping the participants. Finally, we apply the k-means clustering with appropriate number of clusters for extracting groups of similar participants. Using number of clusters equal to 4, we obtain

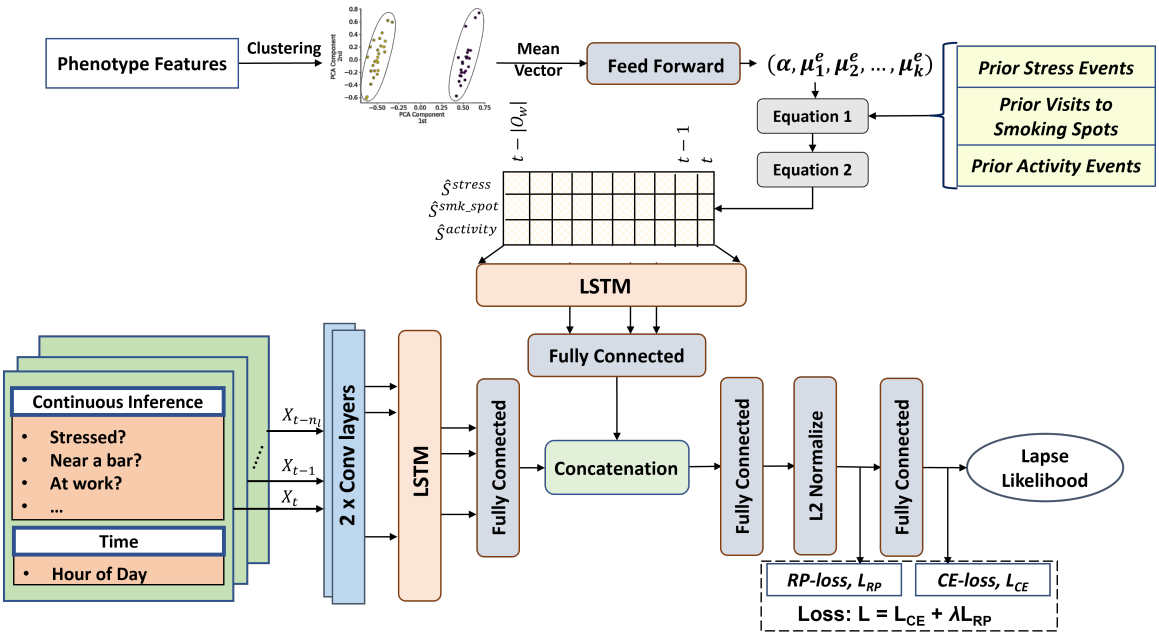


Fig. 3.4: Architecture of the Deep Model with Decaying Historical Influence (DDHI) that uses an explicit model of decaying influence of past events that are expected to wane over time

the best result with all the features contributing positively. The centroid of each cluster is used to estimate the parameters α, $[\mu_1, \mu_1, ..., \mu_k]$.

DDHI Model Architecture

Figure 3.4 shows the overall architecture of the end to end DDHI model. The phenotype features are first used for clustering the participants. The mean of each cluster is then used to output three global context specific parameters (α, μ_1, μ_2) for each event type using a feed forward layer. The centroid of each cluster represents all the participants belonging to that cluster. The centroid is passed as an input through an intermediate feed forward layer. α, $[\mu_1, \mu_1, ..., \mu_k]$ are weights of nodes with sigmoid activation function which are fully connected to the mentioned intermediate layer. Using the appropriate parameters for each event type, the event log in the memory are then transformed to form the event encoded time-series $\hat{S}^{stress}$, $\hat{S}^{activity}$, and $\hat{S}^{smk\_spot}$. Let the length of the current observation window be equal to w with rightmost time t. Then, we output the event of influence encoded time-series for r separate event types

within the current observation window as $\hat{S}^{event_1,\ event_2,\ ...,\ event_r}_{t-w:t} \in \mathcal{R}^{w \times r}$, where $\hat{s}^{stress}_{t-w:t}$ measures the aggregate effect of all past stress events in the current time window $t-w$ to t. The features from continuous inference streams along with the *hour of day* are used in a lagged fashion with multiple observations of $n_f = 47$ features. With n_l such lags, a single instance of lagged features is $X_{t-n_l:t} \in \mathcal{R}^{n_l \times n_f}$. Two separate LSTM networks are trained on top of the lagged features $X_{t-n_l:t}$ and $\hat{S}^{stress,\ activity,\ ...,\ smk\_spot}_{t-w:t}$. We flatten the outputs of LSTM into planar nodes, concatenate the two separate representations and feed it to a multi-layer feed-forward neural network.

3.6 Learning From Sparse & Positive only Labels

Our goal is to estimate the risk of a smoking lapse during the abstinence period from continuous sensor data in the natural environment. We segment the sensor streams using sliding (by 1 minute) *candidate windows* consisting of the observation, intervention, and prediction windows. We assign a positive-label (*high-risk* of lapse) to observation windows only if the corresponding prediction windows overlap with a smoking lapse time, otherwise, the observation windows are unlabeled. Recall that we only consider a lapse to have occurred if it is detected by puffMarker and supported by an EMA. Using either of them alone is insufficient since self-report does not pinpoint the accurate timing of smoking lapse, and puffMarker can produce false positives. As a consequence, our available ground truth labels are sparse, and we only have positive (high-risk) labels available.

3.6.1 Positive Unlabeled (*PU*) Learning

As we only have access to a subset of positively-labeled data and a larger class of unlabeled data which may consist of many lapse events that were either missed by puffMarker, missed by EMA, or missed by both, we adapt positive-unlabeled (abbreviated as *PU*) learning methods to train the *mRisk* model choices. *PU* learning [132] is a variant of the classical supervised learning setup where the assumption is that the data contains positive-labeled or unlabeled samples, which may contain positive (*high-risk* of

lapse) or negative (*low-risk* of lapse) samples. We employ class-weighted base estimators in the *PU* learning framework to address the class imbalance.

As we mark an observation window with a positive label if the corresponding prediction window overlaps with the smoking lapse time, the traditional assumption that positively-labeled data is *selected completely at random* (SCAR)) does not hold. Therefore, we use the *PU-bagging* or ensemble *PU* learning approach [63] that is independent of the SCAR assumption and use *leave-one-participant-out-cross-validation (LOPOCV)*.

Researchers previously employed PU learning methodology to train classical machine learning models for estimating the risk of smoking lapse [24]. Adopting a similar approach, we train deep neural networks using the PU Bagging approach of model learning. We propose a novel loss function, Rare Positive Loss to train our models. The evaluation section (see Table 3.1) documents the gain in performance using our proposed model training approaches.

3.6.2 Rare-Positive (*RP*) Loss Function

Key to training deep learning models is a suitable loss function that the model can use to optimize the representation. Contrary to the typical supervised learning setup, where concrete ground truths are available for both positive and negative classes, we only have access to a subset labeled positives (i.e., *high-risk* moments). All the other samples are unlabeled and consist of positives (i.e., lapses missed by puffMarker and/or EMA) and negatives (*low-risk* moments); we assume that the proportion of positive instances is rare in the unlabeled class. We want to guide the learning process so that the model learns an accurate representation of the positive class and learns to extract other rare true positives from the unlabeled class.

Design of the RP Loss Function

In designing the *RP* loss function, we aim to achieve two key goals. First, we want to create a representational feature space in which positive data points are

clustered together. This is trivial for the model to do by coalescing all the input instances into a single point in the feature space. Hence, the second condition needs to be designed, which constraints such development. Our second competing goal is to ensure that the learned representation space of the positive class can only include a small portion of the unlabeled class, as positive instances are expected to be a rare occurrence in the unlabeled class. To formulate the two components of our proposed loss function, we let $\mathbb{S}$ denote the set of all samples, $\mathbb{S}_p$ the set of all positively-labeled samples s_p and $\mathbb{S}_u$ the set of all unlabeled samples s_u, with $\mathbb{S} = \mathbb{S}_p \cup \mathbb{S}_u$.

Positive Class Dispersion (P): We adopt the definition of consistency as proposed recently in [169], to minimize intra-class variations, but apply it only to the positive class ($\mathbb{S}_p$). Our goal is to reduce the mutual dispersion of the positive instances for forming dense clusters. As in [169], our data is also collected by wearables in the noisy field environment, and hence are impacted by outliers. To reduce the impact of outliers, we also define dispersion of the positive class ($\mathbb{S}_p$) in terms of a robust aggregate function.

Consistency of $s_p^i \in \mathbb{S}_p$ is the average distance of its representation from the representation of all other points $s_p^j \in \mathbb{S}_p, i \neq j$, in the model's feature space, i.e., $C(s_p^i) = d(s_p^i, \mathbb{S}_p)$, using the definition of average distance in the feature space from [169]. It was shown in [169] that this definition of distance is differentiable and hence suitable for use in loss function and leads to faster convergence (for noisy data collected by wearable devices). Now, consistency of the positive class is defined as an aggregated function, ψ, of all the point consistencies within the class. Within a mini-batch of data $\mathbb{U}_{MB} \in \mathbb{S}$, positive class dispersion, P is defined as

$$P = \psi\left(\{C(s_p^i)\}_{s_p^i \in \mathbb{U}_{MB} \cap \mathbb{S}_p}\right) \tag{3.3}$$

Similar to [169], we also select a percentile measure for ψ. But, in contrast with [169]

that uses non-overlapping windows of data, we need to produce a risk score for each minute and hence use overlapping windows, sliding each minute. Consequently a positive event (i.e., a confirmed lapse) is contained in all overlapping observation windows whose prediction window (e.g., 60 minutes long) contains the positive event. One positive event a day can result in 10% (60 out of 600 minutes of sensor wearing a day) of the data labeled as *high-risk*. Therefore, we use 80^{th} percentile of the point consistency values of the positive class to obtain robustness, while respecting rarity of the positive class. Minimizing P ensures that the positive instances pack tightly in the deep representations space.

Rarity of the Unknown Positives Within Unlabeled Class (R): Given the assumption of rarity of positive samples in the unlabeled class, the tight cluster produced for the positive class (by minimizing P) should only contain a small portion of the unlabeled class. For this purpose, we define the rarity metric R as the proportion of unlabeled samples whose average distance from the samples of positive class are at most P.

Let $d(s_u^i, \mathbb{S}_p)$ denote the average distance of the representation of unlabeled sample $s_u^i \in \mathbb{S}_u$ from the representation of all positive instances $s_p^j \in \mathbb{S}_p$ in the model's feature space. We define an indicator function

$$I(s_u^i) = \begin{cases} 1 & d(s_u^i, \mathbb{S}_p)) \leq P \\ 0 & \text{otherwise.} \end{cases}$$

Our goal is to limit the number of unlabeled instances for whom the above indicator function outputs 1. For this purpose, given a mini-batch of data instances $\mathbb{U}_{MB} \in \mathbb{S}$, we define the rarity metric R as follows.

$$R = \frac{\sum_{s_u^i \in \mathbb{U}_{MB}} I(s_u^i)}{|\mathbb{U}_{MB} \cap \mathbb{S}_u|} \tag{3.4}$$

Minimizing R amounts to reducing the proportion of unlabeled instances which fall within the cluster of positive instances and minimizing P constraints the positive instances to form a tight cluster itself.

We compose our overall *Rare-Positive (RP)* loss function as follows so the model can concurrently optimize both positive dispersion (P) and rarity (R) measures.

$$\mathcal{L}_{RP} = \gamma P + (R - \epsilon)^2, \tag{3.5}$$

where ϵ is the expected proportion of unknown positives we assume to be present within the unlabeled class. $(R - \epsilon)^2$ denotes the squared distance of the rarity metric R from a fixed ϵ value. We choose the quadratic function in favor of an absolute error for two reasons. First, quadratic error term is continuously differentiable. Second, we want the penalty for an error to increase in proportion to the magnitude of the error itself.

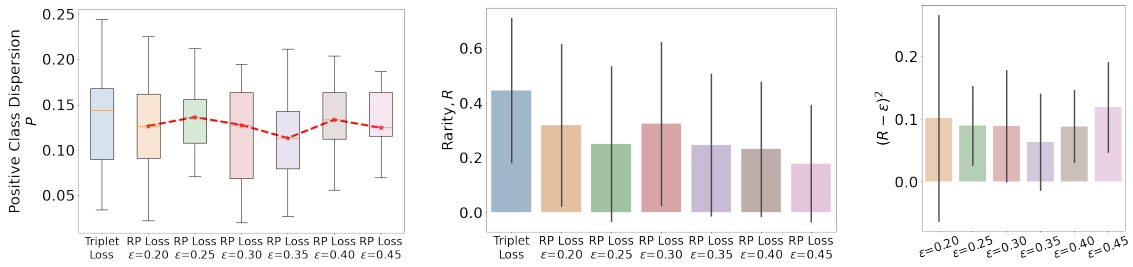
We conduct experiments to find the best value of ϵ from our dataset. The γ value is a scaling hyper-parameter for scaling two terms with different units. Since, we L_2 normalize the deep vectors to have unit norms before distance calculation, their range is similar to the range of proportions $(0, 1)$. We choose $\gamma = 0.5$ for our experiments.

The Loss Function

For training the *mRisk* model, we employ the joint supervision of *cross-entropy* loss (to derive risk likelihood between 0 and 1) and *RP* loss. More specifically, our overall loss objective is

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{RP}, \tag{3.6}$$

where $\mathcal{L}_{CE}$ is cross-entropy soft-max loss [170] and we use λ $(= 0.2)$ to balance the effect of two loss functions.



(a) (b) (c)

Fig. 3.5: P and R values when using different values of ϵ in the *RP* loss function, compared with that from using Triplet loss

3.7 Optimization, Evaluation, and Explanations of *mRisk* Model Choices

We first determine the best value of the hyper-parameter ϵ to optimize the proposed *RP* loss function. Second, we compare the performance of our two proposed models by analyzing the risk characteristics each model produces. Third, we design simulation experiments to evaluate how successful the models are in creating intervention opportunities prior to each confirmed lapse. Fourth, we visually analyze the risk dynamics produced by *mRisk* before and after lapse moments. Finally, to understand the major factors driving the lapse risk produced by the *mRisk* model, we explain the influence of features on the model performance using Shapley values [151].

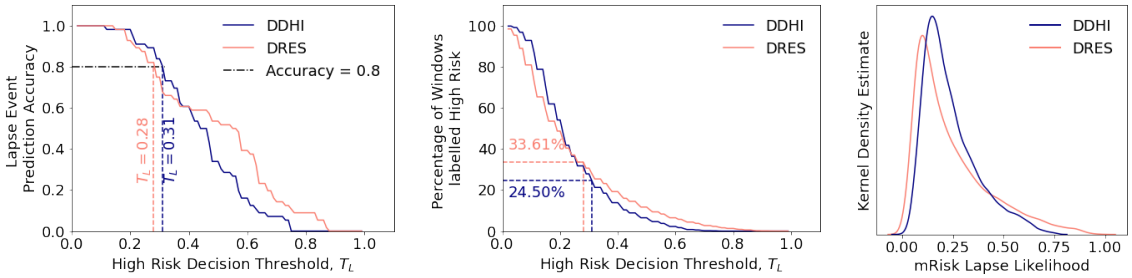
3.7.1 Loss Function Optimization and Evaluation

We experiment with different choices of ϵ (which denotes the expected proportion of rare positives within the unlabeled class) on positive class dispersion (P) and rarity metric (R) to determine its best value. We also compare with Triplet loss [171], a widely used traditional loss function used in deep learning. Figure 3.5 shows the results when we train the models by combining the stated loss functions with cross-entropy loss. We make several observations. As each model is trained with mini-batches, we first analyze the distribution of P and R for different choices of ϵ. We observe that the model achieves lowest deviations (or spread) in P and $(R-\epsilon)^2$ for $\epsilon = 0.35$. We take this as an indication that for this value of ϵ, the model is able to

consistently find the best representation to separate out positives (including unknown positives in the unlabeled class) from the negatives (all in the unlabeled class). We get another supporting indication of it by observing that the value of P is the lowest for this choice of ϵ. We see from Figure 3.5c that when ϵ increases from 0.2 to 0.35, the weight assigned to the $(R-\epsilon)^2$ component of the *RP* loss function reduces because $0 \leq (R-\epsilon) \leq 1$. After this value, ϵ gets farther away from the true proportion of positives in the unlabeled class (see Figure 3.5b), making it harder for the model to find a suitable representation. Therefore, we hypothesize that for $\epsilon = 0.35$, the model is able to find a representation to form the tightest cluster of positives while allowing unlabeled positives. We use $\epsilon = 0.35$ for all experiments.

We next observe from Figure 3.5b that at $\epsilon = 0.35$, the proportion of unlabeled positives is 24.68% of the unlabeled data (i.e., R). We use EMA reported lapses that were not used in model training (as they were missed by puffMarker) to estimate the proportion of positive class in unlabeled data. Each EMA where one or more lapses was reported, indicates a 2-hour lapse window where participants recall having smoked. If these hours are considered to represent high-risk moments, they constitute 17.8% of all unlabeled hours of data. As the high-risk moment is considered to precede a smoking lapse, the entire 2-hour window may not constitute high-risk moments, while hours where no lapse was reported may also consist of high-risk moments, this is only a crude estimate based on available sources of imprecise labels. Nevertheless, the two estimates lie within 7% of each other.

We also observe that treating the unlabeled data as negatively labeled and using Triplet loss to maximize its separation from the positive class results in a representation that produces slightly higher values of P as the *RP* loss function (especially for $\epsilon = 0.35$). But, as the model is forced to maximally separate positives from the unlabeled class, it ends up admitting a larger proportion of unlabeled data (about 45%) in the positive cluster. Using a model trained with such a loss function will require a



(a) Behavior of lapse detection accuracy.

(b) Proportion of samples labeled *high-risk*.

(c) Distribution of risk scores.

Fig. 3.6: Evaluating *mRisk* model choices on *PU-labeled* data

higher number of interventions to achieve a given recall rate (i.e., intervention delivered prior to a detected lapse event) as compared with the *RP* loss function.

3.7.2 Evaluating *mRisk* Model Choices by Their Risk Characteristics

We train the two *mRisk* model alternatives using only sparse positive labels. Lack of unambiguous negative labels of *low-risk* moments diminishes our options of computing traditional metrics such as *F1 score*, *AUROC*, and others. Thus, we opt for measuring the performance of the models in predicting the detected lapse events. If the model outputs a *high-risk* probability for a confirmed smoking lapse, we consider it an accurate prediction.

However, if we classify every data-point as *high-risk*, we would achieve 100% accuracy. In a traditional supervised learning setup, we measure the false positives, which gives us a measure of the cost of using/deploying any developed model. Since we can not measure the false positive rate directly, we propose to measure the cost of our model indirectly. At every decision point, we treat the percentage of assessment windows determined to be *high-risk* as the cost of a specific model. This indirectly captures the user burden posed by a model in real-life where a *high-risk* moment may trigger an intervention to reduce the likelihood of a lapse.

We also note that considering any data-point as *high-risk* requires specifying a decision threshold (T_L) within the probability scale. If the model outputs a probability

$\geq T_L$, we consider it a *high-risk* moment, and *low-risk,* otherwise. We select a value of T_L to achieve a lapse detection accuracy of 80% and report the inference cost.

Results

Figures 3.6a and 3.6b together captures the trade-off between lapse detection performance and the inference cost for using different values of the decision threshold (T_L). Figure 3.6a shows the steep drop-off in detection accuracy for both the *mRisk* model choices as we increase the value of the decision threshold. The drop-off in accuracy is comparatively less steep for the *DDHI* model when compared to the *DRES*. For achieving a minimum of 80% lapse detection performance, the decision threshold values are 0.28 for the *DRES* model and 0.31 for *DDHI*. The corresponding inference costs are 33.61% and 24.50% for *DRES* and *DDHI* respectively. Thus, for the same lapse detection performance, we obtain a 9% improvement in the inference cost by using the *DDHI* model. Figure 3.6c shows the distribution of the lapse likelihoods produced by both models. Both models have the desirable right-skewed distribution, as we expect a majority of moments to represent a low risk.

3.7.3 Evaluating *mRisk* Model Choices via Simulated Delivery of Risk-Triggered Interventions

For our next evaluation of the two models, we train a baseline machine learning model and evaluate how successful the models are in creating intervention opportunities prior to each confirmed lapse. We design simple simulation experiment where interventions are delivered when the risk for lapse rises above a pre-determined threshold (T_L) (see Section 3.7.2). To limit intervention fatigue [172], no new interventions are triggered until *intervention gap (I_G)* minutes have elapsed since the last intervention, assuming the impact of an intervention lasts at least this long.

Since we use a prediction window of 60 minutes, we use $I_G = 60$ minutes. We note that introducing an intervention gap changes the direct relationship between the risk threshold and the frequency of interventions observed in Section 3.7.2. Although the

choice of T_L and I_G will depend on the characteristics of the dataset, preferences of the smoking intervention researcher, and other real-life constraints (e.g., no intervention when driving or when in meetings), we analyze the performance of the *mRisk* model choices in the simple scenario when the intervention delivery only depends on T_L and I_G to show its expected behavior. Keeping I_G set at 60 minutes, we vary T_L to observe the performance of each model at different frequency of interventions per day.

Evaluation Metric

For each model, we estimate the probability that an intervention opportunity is available ahead of a lapse event. For this purpose, we use only the confirmed lapse moments, i.e., positive labels. The proportion of lapse events occurring within 60 minutes (i.e., prediction window) of an intervention is called the *Intervention Hit Rate (IHR)*

Intervention Hit Rate (*IHR*) measures the probability that an intervention opportunity is provided by *mRisk* ahead of each lapse event, i.e., within our prediction window (P_w). More formally, we first choose a value for risk threshold, $T_L = c$ to achieve a desired frequency of interventions per day. An intervention opportunity at time t is created if the risk produced by *mRisk*, $r(t)$ exceeds c and at least I_G (intervention gap) minutes have elapsed from the most recent intervention moment. Let $I(u) = \{t_1^i(u), t_2^i(u), t_3^i(u), ...\}$ be the set containing the timings of interventions for a user (u). Let $L(u) = \{t_1^l(u), t_2^l(u), t_3^l(u), ...\}$ be the precise time of a lapse events for user u (confirmed by EMA and Puffmarker). We consider the lapse event at time $t_k^l(u)$ to be intervened (or covered or *hit*) if $\exists j : t_j^i(u) \leq t_k^l(u) : t_k^l(u) - t_j^i(u) \leq P_w$. The IHR can then be defined as

$$\frac{\sum_u |\forall k : \exists j : t_j^i(u) \leq t_k^l(u) : t_k^l(u) - t_j^i(u) \leq P_w|}{\sum_u |L(u)|}.$$

We note that $t_j^i - t_{j+1}^i > I_G$, i.e., no successive interventions are at least I_G minutes apart. Therefore, there exists a unique j for each lapse moment, if $I_G \geq P_w$.

As launching an intervention at every allowable moment can trivially achieve a 100% *IHR*, but at the cost of a high intervention frequency, we measure the participant burden via intervention frequency and determine *IHR* for different values of intervention frequency per day. For a given intervention frequency, a better model should have a higher *IHR*.

Experiment Setup

We simulate with an intervention frequency range of $[3, 7]$ per waking day to evaluate *mRisk* model choices — *DRES* and *DDHI* — in creating intervention opportunities. We also train a Random Forest Model using the PU Bagging Framework, named *PU-Bagging RF* [24], to act as a baseline. This model accepts the feature vector used in the *DRES* model, and produces a risk score for each observation window.

To vary the intervention frequency per day for the *PU-Bagging RF*, *DRES* and *DDHI* models, we vary the risk thresholds. In addition to evaluating the performance of the three models on IHR, we also compare the difference in IHR when using the new RP loss function versus Triplet Loss in both *mRisk* model choices. To evaluate the impact of phenotyping idea in the *DDHI* model, we experiment with different number of phenotypes, including no phenotypes. Finally, as learning the personal smoking spots for each new user requires collecting and analyzing pre-quit data, we evaluate the gain in performance when this data is used in modeling.

Table 3.1: Intervention Hit Rate at Different Frequencies of Intervention for Different Models

		IHR at Different Frequencies of Intervention								
Model	**Loss Function**	**3**	**3.5**	**4**	**4.5**	**5**	**5.5**	**6**	**7**	**Mean IHR**
PU-Bagging RF [24]	—	0.30	0.37	0.49	0.64	0.70	0.75	0.75	0.76	0.60
DRES	Triplet loss	0.44	0.51	0.57	0.68	0.74	0.78	0.84	**0.93**	0.69
DRES	RP loss	0.46	0.55	0.64	0.74	0.76	0.78	0.84	**0.93**	0.71
DDHI	Triplet Loss	**0.51**	0.59	0.65	0.71	0.73	0.80	0.85	0.86	0.71
DDHI	RP loss	0.50	**0.62**	**0.68**	**0.74**	**0.76**	**0.85**	**0.89**	**0.93**	**0.74**
DDHI Without Personal Smk. Spots	RP loss	0.47	0.51	0.55	0.60	0.66	0.75	0.80	0.91	0.66

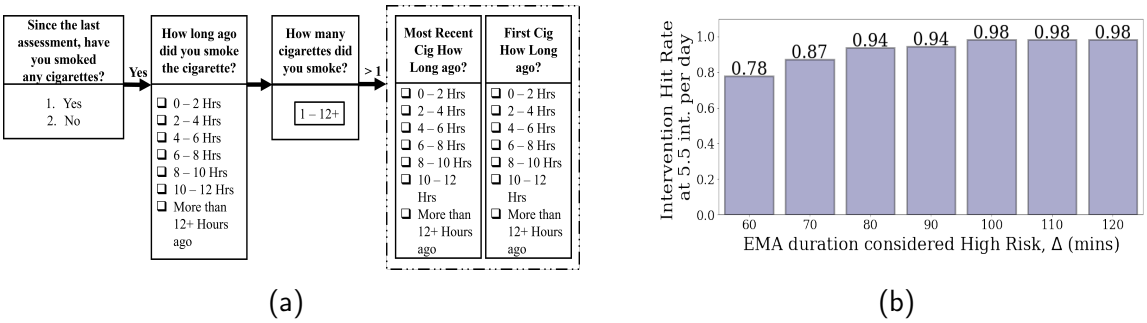
Results

Table 3.1 shows that *DRES* and *DDHI* outperform the baseline *PU-Bagging RF* model, *DDHI* outperforms *IHR*, and RP Loss outperforms Triplet Loss. The last row in Table 3.1 shows that not using personal smoking spots results in a substantial drop in performance of both models.

Table 3.2: Intervention Hit Rates obtained from *DDHI* model with different number of phenotypes

	IHR at Different Frequencies of Intervention								
No. of Phenotypes	**3**	**3.5**	**4**	**4.5**	**5**	**5.5**	**6**	**7**	Mean IHR
No Phenotyping	**0.53**	0.56	0.59	0.70	0.74	0.77	0.88	**0.93**	0.71
2	0.52	0.58	0.66	0.70	0.71	0.81	**0.89**	**0.93**	0.73
4	0.50	**0.62**	**0.68**	**0.74**	**0.76**	**0.85**	**0.89**	**0.93**	**0.74**
6	0.51	**0.62**	0.64	0.71	**0.76**	0.81	0.87	**0.93**	0.73
8	0.52	0.60	0.65	0.71	0.75	0.81	0.83	**0.93**	0.73

Table 3.2 shows the effect of phenotyping in the *DDHI* model. We observe that increasing the number of phenotypes improves *IHR*, achieving a peak IHR for four (4) phenotypes suggesting it as the optimal for our dataset. As the DDHI model with RP Loss function outperforms other models, we select this as the *mRisk* model in subsequent experiments. We select 5.5 interventions per day, as it provides the largest jump in IHR. We also find that for this choice, the risk crosses the threshold approximately 32 minutes prior to the lapse moment, on average, providing half an hour window to intervene prior to a lapse.



(a) (b)

Fig. 3.7: (a) Shows the EMA items corresponding to smoking report by individuals, (b) Intervention Hit Rate at 5.5 int. per day when considering a certain duration of EMA response as positive lapse

3.7.4 Evaluating *mRisk* Model Performance on Training-Independent EMA Labels

In the preceding evaluation (in Section 3.7.3), we only used those lapses reported in EMAs that was also detected by Puffmarker providing us with precise moment of lapse, in estimating the intervention hit rate (IHR). These labels were also used in the model training. For a more independent evaluation of the *mRisk* model, we use a new source of lapse labels from our field data that was not used in training or testing of the model. These are lapses reported in EMA's that were missed by puffMarker (due to lack of sensor data or model failure). Figure 3.7a shows an EMA that participants fill out to report recent cigarette smoking lapses. If users report smoking, they are asked to report the time of smoking and the amount of cigarettes they have smoked. If they report smoking more than one cigarette, they are also asked to report the timing of the first and most recent cigarette. We use three questions related to reporting the time of smoking events — 'How long ago have you smoked?', 'How long ago you smoked first cig', and 'Most Recent cig how long ago?'.

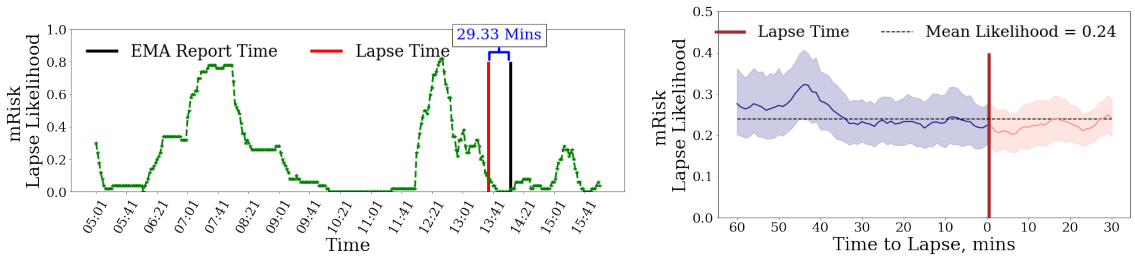
As Figure 3.7a shows, participants indicate a 2-hour time window. When an EMA report of lapse is missed by puffMarker, we are unable to determine the precise moment of lapse and can only locate it in a 2-hour *lapse window*. Therefore, these labels are not used in training the models. In the absence of precise lapse moment, we

consider the entire lapse window as the potential lapse time. For example, if at time t, a participant reports smoking a cigarette '4 - 6 Hours' ago, we label $t-6$ Hours to $t-4$ Hours as containing a smoking event. The actual lapse event may occur anywhere in a specific lapse window, and hence the high-risk moments (that are assumed to precede a lapse) may occur at different portions of the 2-hour lapse window.

We adopt the following approach for computing the intervention hit rate for EMA-reported lapses. Let t_{int} denote the time when the estimated risk produced by the pre-trained *mRisk* model crosses a prespecified threshold (corresponding to an expected 5.5 interventions per day) and triggers an intervention. Let $[t_{EMA}, t_{EMA}+2H]$ denote the lapse window based on the participant's EMA response. We say that the intervention delivered at time t_{int} has preceded a lapse if the prediction window $[t_{int}, t_{int}+P_w]$ has an overlap with $[t_{EMA}, t_{EMA}+\Delta]$. Here, Δ denotes the duration of time since the start of the 2-hour lapse window considered as high risk. If $\Delta = 60$ minutes, then only the first hour of the 2-hour lapse window is considered to be high-risk. If $\Delta = 120$ minutes, then the entire lapse window is considered high-risk. We assume that risk is high prior to a lapse and low afterwards, which is confirmed by our subsequent analysis (see Section 3.7.5).

We use 2-hour lapse windows that have risk scores available from the *mRisk* model at least 30 minutes (depending upon the availability of sensor data, including imputed data for short periods of missing sensor data). This results in a total of 615 lapse windows reported in 336 EMA's that are used in this analysis.

We vary the value of Δ from 60 to 120 minutes and report the intervention hit rate in Figure 3.7b corresponding to 5.5 interventions per day. We observe that IHR increases from 0.78 and saturates at 0.98 for $\Delta = 100$ minutes, indicating that most high-risk moments are contained within the first 100 minutes of the 2-hour lapse window. As the actual lapse moment and the actual high-risk moment may vary from instance to instance, the IHR reported here may represent an overestimation. Nevertheless, this



(a) (b)

Fig. 3.8: Lapse Likelihood produced by the *DDHI* model with lapse, intervention and EMA report times shown with vertical lines. We only include those EMAs in which the participants confirmed that the last time they smoked was 0-2 hours ago.

analysis shows that the *mRisk* model may enable the delivery of an intervention prior to most self-reported lapses, even at the rate of 5.5 interventions per day.

3.7.5 Rise/Fall in Risk Levels Produced by *mRisk* Before/After Lapse Moments

As the *mRisk* model produces a continuous risk score, we visually analyze the rise and fall in the risk scores before and after lapse moments. We first apply the *mRisk* model post-facto on daylong data from a participant in Figure 3.8a. The moment of lapse from puffMarker is shown together with the time when the accompanying self-report of lapse was recorded. We make several observations.

First, we observe that for the case when both detected and reported lapse are available (see Figure 3.8a), the reported time is 29.33 minutes after the actual lapse in this instance. In other instances, this time gap may be higher or lower. This ambiguity in determining the actual timing of lapse makes it difficult to use self-reported lapses (not supported by sensor-based detection) for model training or testing.

Second, in Figure 3.8a the lapse is preceded by a high-risk episode as estimated by the *mRisk* model. We further observe that as time gets closer to the lapse moment, the risk decreases. We also observe that once lapse occurs, the risk falls further, perhaps due to satiation of smoking urge.

Third, we observe two *high-risk* windows in the entire day. The *mRisk* model can

guide the delivery of an intervention prior to the risk reaching its peak during both the *high-risk* episodes.

Figure 3.8a only shows the variation in risk score around one lapse moment for a single participant. To see if there is a general pattern of risk rising prior to lapse and falling immediately before and after the lapse moment, we aggregate the risk scores across all lapse moments from all participants. Figure 3.8b shows the mean lapse risk (with a confidence interval of 90%) before and after a smoking lapse. The mean risk score is also plotted. We observe that generally, the risk score is around the mean level. But, it rises and peaks around 44 minutes prior to a smoking lapse. The risk then decreases as the time approaches the lapse moment, falling below the mean level at the time of lapse, and falling even further after the lapse moment. We note that even though the observed variability may diminish when data from different lapse instances are pooled, due to the risk peaking at different times for different lapse instances, we still see a robust pattern at the population scale.

3.7.6 Understanding the Role of Context in Estimating Lapse Risk via Model Explanations

For the *mRisk* model to be trusted by intervention researchers [173], we analyze the behavior of the *mRisk* model in terms of the influence of the three major sensor-derived contexts (i.e., stress, activity, and location) on the lapse risk. We utilize the *SHapley Additive exPlanations (SHAP)*, a game theory-based algorithm that can be employed to explain global and local feature importance for a fitted machine learning model [151]. *SHAP* explains a prediction by assuming that each feature value of the instance is a *player* in a game and the final prediction is a *payout*. Based on coalition game theory principles, the algorithm assigns payouts to players depending upon their contribution to the total payout. Players cooperate in the coalition and receive specific *profits*. In our case, the *payout* is the prediction of the risk of lapse for a single instance. The *profit* is the actual prediction for this instance minus the average prediction across

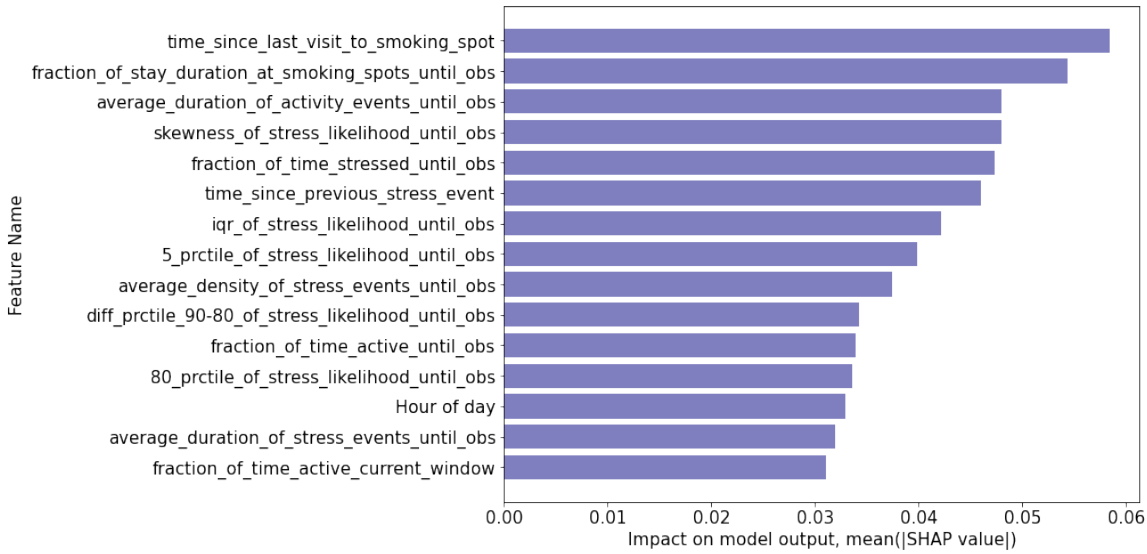


Fig. 3.9: Global Feature Importance showing top 10 features for *DRES* model using *Deep SHAP*

all instances. The *Shapley* value is the weighted marginal contribution of a feature across all the possible coalitions. Features with large absolute *Shapley* values are more important.

We approximate the *Shapley* values for each input node of the *DRES* using the *Deep SHAP* method proposed in [151]. *Deep SHAP* builds upon DeepLIFT [174], which is a local additive feature attribution method for approximating the conditional expectations of *SHAP* values using a collection of background samples (training data, see [175] for details). Using *Deep SHAP*, we first obtain the *Shapley* values of each input instance ($n_t \times n_f$, $n_f = 62$) of the *mRisk* model. We then average the *Shapley* values of each feature along the time axis. Finally, *Shapley* values of all instances across all participants are aggregated to interpret the collective impact of the input features on the model (i.e., *global feature importance*).

Observations from Global Feature Importance

Figure 3.9 shows the impact of top *15* features on the *mRisk* model output, ranked by their *Shapley* values, averaged over all iterations. The top features are

distributed across multiple contexts — visiting smoking spots, stress, activity, and hour of day. The most influential feature (*time since last visit to smoking spot* and *fraction of stay duration at smoking spots until obs*) indicate that exposures to smoking spots influences lapse risk. *Average duration of activity events until obs* also has significant influence. We hypothesize that spending more time moving around increases the chance of exposure to environmental cues of smoking, which may increase the risk of lapse.

We observe that 9 out of the 15 features are related to stress. These include *skewness of stress likelihood until obs*, *fraction of time stressed until obs*, and *average density of stress events until obs*. We hypothesize that frequency and duration of high-stress likelihood so far in the day influences the risk of lapse. We also observe the event-of-influence features, which encode the temporal dynamics of recent contexts, outrank the continuous inference features. This observation underscores the importance of suitably representing the events-of-influence time series in a deep modeling framework that utilizes these contexts for learning.

3.8 Discussion, Limitations, and Future Works

Although this work uses a specific application of smoking lapse and a specific real-world dataset, the many interesting challenges encountered in modeling and the proposed ideas to address them may be applicable in the continuous estimation of risk in related domains such as the risk of lapse when quitting excessive drinking, abstaining from addictive substances (e.g., cocaine), controlling overeating, overcoming suicide attempts, among others. Like smoking, each of these adverse behaviors occurs in the natural environment. Similar to smoking lapses, they are influenced by both internal states and external cues. Mobile sensor data can passively track risk factors for each of these, but they are likely to be similarly noisy. Finally, the timing of a subset of adverse events may be obtained, but getting unambiguous negative labels is similarly difficult.

The *mRisk* model proposes a new end-to-end framework for model development that may be adaptable to continuously estimate the risk of other adverse behaviors. It

presents approaches to incorporate the influence of both recent and past events captured from imperfect machine learning models applied to noisy sensor data and proposes a new loss function with customizable parameters to train a model for continuous risk estimation. It also proposes approaches for evaluating modeling choices in the absence of unambiguous negative labels by using the limiting of intervention burden in place of negative models to guide the model optimization. It also shows an approach for evaluating the expected utility of such risk models in a simulated delivery of interventions.

3.8.1 Key New Insights

For estimating the risk for smoking lapse in newly abstinent smokers, the *mRisk* model led to several new insights. First, it helped determine the proportion of unlabeled data that is likely to represent a *high-risk*. Second, we find that determining the personal smoking spots during the pre-quit period and using them in risk estimation can lead to substantial improvement in the model performance. Third, via visual analysis of the continuous risk estimates produced by the *mRisk* model, we find that lapse risk peaks about 44 minutes prior to an impending lapse, providing sufficient opportunity to intervene. Fourth, we find that 85% of lapses can potentially be intervened upon with only 5.5 interventions per day. Finally, via explanation, we find that recent exposure to smoking spots has a large influence on the lapse risk together with being physically active and a high likelihood of recent stress.

3.8.2 Limitations and Future Works

This work is only a first step towards continuous estimation of risk for adverse behaviors using mobile sensors that can be used in real-life field settings. It has several limitations that present exciting opportunities for future research for both computing and health researchers. First, many smoking lapses captured in EMAs could not be used in our model development or evaluation as they were not detected by puffMarker, preventing a precise determination of the time of lapse. The EMAs locate the past

smoking events (sometimes more than one) within a 2-hour long window. This does not allow a determination of which segments of sensor data within this 2-hour window correspond to moments prior to a lapse and can be labeled high-risk. Future work can explore novel ideas to make use of these temporally-imprecise label sources to further improve the model.

Second, future work can also explore ways to identify moments of low risk via EMA responses and use them to train the usual two-class models. Third, this work shows the direct applicability of the presented *mRisk* framework to estimate the risk of smoking lapse. Applying it to other datasets of smoking cessation may require adaptation of some parameters such as the ϵ value in the *RP* loss function and the choice of percentiles in deciding the value of P. Future work can explore how well the *mRisk* framework may be used to estimate the risk of other adverse behaviors (e.g., alcoholism, drug addiction, etc.) that also have noisy data and incomplete and positive-only labels.

Fourth, the *mRisk* model achieves a good recall (*IHR*) using only the stress, location, and activity features. Future work can boost the performance further by supplementing them with craving, self-efficacy, presence of other cues such as noisy locations, graffiti, and other situational indicators that may affect the risk of lapse. Another idea to improve the model performance may be to use self-report data from EMAs in the context of research studies that collect EMAs for other purposes. Fifth, our simulation of intervention delivery only uses an intervention gap to avoid fatigue from frequent interventions. Future work can improve its real-life applicability by incorporating other constraints such as users' receptivity [176] and availability [177].

Sixth, our evaluations assume that interventions can be delivered as soon as *high-risk* moments are detected if permitted by other constraints. But, how the detection of *high-risk* moments can be used to deliver the most efficacious intervention requires a just-in-time-adaptive-intervention (JITAI) optimization trials (e.g., micro-randomized trial) [148] to determine the best conditions (e.g., *high-risk*, moderate

risk, or *low-risk*) and the best corresponding combination of the intervention content, mode of delivery, and the adaptation mechanisms for personalizing the intervention to the individual based on his/her contexts. Seventh, risk scores produced by *mRisk* can potentially be used to evaluate the impact of interventions that target stress reduction, location exposure via geofences, nicotine medications, and others in reducing the lapse risk. Eight, the risk scores along with the driving factors can be presented to newly abstinent smokers at the end of the day to help them understand their vulnerabilities better. Finally, *mRisk* is an offline model, computed only from observational data after data has already been collected. However, to be widely used for sensor-triggered mobile intervention during micro-randomized trials, future work can implement an online version of the *mRisk* model to run on wearable devices or smartphones. Only then can the model be used to trigger real-time mobile interventions based on the online prediction of the risk of a lapse in the natural environment of the participants. These make for exciting future research agenda for the computing and health research community.

3.9 Chapter Summary

The majority of chronic diseases can be prevented or better managed by improving health-related behaviors. Automated detection of risky contexts via mobile (and wearable) devices provides a new opportunity to improve the success rate with behavior modification. But, the overall risks depend on a multitude of factors, including internal states, personal behaviors, and environmental cues. Many of these factors can now be detected by applying machine learning models on data collected by wearable devices and smartphones. But, the challenge is noise in the data collected and lack of unambiguous labels of low- and high-risk moments. This work provides a new framework to estimate the overall risk of adverse behaviors despite noisy data, no labels of low-risk states, and availability of only a subset of high-risk states. It shows the successful application of this model on smoking cessation dataset, opening the doors for exciting

new opportunities in the design and delivery of efficacious behavioral interventions to help people live healthier lives.

Chapter 4

Robust Inference of Human States from More Convenient, but Noisier Wrist Sensor Data

4.1 Introduction

In Chapter 3, we developed "mRisk," - a continuous smoking risk estimation model using mobile health sensors in the wild. We employed the chest-based AutoSense [72] sensor suite to continuously collect ECG and Inertial Motion data from participants in their natural environment. The ECG data is first processed to compute human heartbeat intervals (RR intervals). Next, we calculate cardiac and heart rate variability features from the RR interval time series with a sliding window approach. Finally, using a machine learning-based model, we transform the feature sets to continuous assessments of participants' physiological stress levels in the natural environment [90, 91, 92, 15, 93, 94]. A variety of wearable ECG devices have also been used to develop continuous detection models for a range of complex activities and behaviors that require cardiac-related features such as the detection of drug use [88, 87, 86], craving [30] and pain [178, 179].

The chest-worn IMUs (Accelerometer and Gyroscope) directly estimate the participants' torso movement, and we use them to compute participants' activity levels. The activity time series represents the behavioral context of the participants. Stress, activity, and proximity to smoking spots (from smartphone GPS) are the inputs to the prediction models. A limitation of the mRisk model development approach presented in Chapter 3 is the use of chest-based wearable sensors to infer these physiological (e.g., stress), behavioral (e.g., activity), and environmental (e.g., proximity to smoking spots) contexts. Platforms like AutoSense [72] achieve reliable attachment and minimal signal noise by using adhesive electrodes that are not comfortable for extended use. It essentially limits the platform to research use cases. Other platforms like the Zephyr Bioharness [180] achieve improved wearability levels by using conductive fabrics instead

of adhesive electrodes. Such devices have been used in athletic performance monitoring and research studies. However, the resulting variable attachment results in lower overall signal quality. In addition, the chest belt form factor is still not practical for long-term, continuous monitoring outside research studies. Therefore mRisk models based on chest-worn sensors offer limited practical utility with reduced potential for successful deployment in smoking cessation programs. To realize the broader potential of the methods and models developed in Chapter 3, we must be able to develop risk prediction models using convenient and easy-to-use devices and sensors. Hence, adapting the mRisk models to work with wrist-based sensors instead of chest-based ones is essential. Enabling continuous smoking risk estimation from wrist-based sensors using mRisk first requires robust inference of participants' stress and activity levels using data collected from wrist-worn sensors alone.

The last decade of research in the ubiquitous computing community has seen a drive towards the continuous detection of increasingly complex activities and behavioral states using devices that are increasingly unobtrusive and more practically deployable, both for use in research studies and everyday life. Decades of research on human activity recognition using wearables IMU sensors have given rise to an established set of methodologies for computing activity levels from participants in their daily lives. Intertial motion sensors fitted in wrist-worn wearables have seen widespread adoption in consumer health analytics owing to their ability for caloric estimation, human activity, and mobility pattern recognition. We borrow from this vast literature to develop a deep neural network-based human activity recognition model using accelerometer data collected from the wrist. However, contrary to activity recognition from the wrist, inferring stress is more complex. Computing heart rate variability and cardiac features from lower-frequency and noisier wrist-worn Photoplethysmography (PPG) is more challenging than chest-based wearable Electrocardiogram (ECG) devices. The Apple Watch (and similar devices) [181] provides an alternative to chest belt-based ECG sensing in a very

different part of the design space that achieves greatly improved usability through a wristband form factor. However, the Apple Watch is optimized to produce short ECG strips suitable for diagnosing arrhythmias. It requires the user to touch the watch's crown (which functions as an electrode) throughout the sampling process - implying the user must interrupt their normal activities to collect ECG samples. As a result, this approach is unsuitable for continuous, ubiquitous monitoring applications. By contrast, wrist-worn fitness trackers and smartwatches that include a Photoplethysmography (PPG) sensor (e.g., FitBit and Garmin fitness trackers, WearOS watches, the Apple Watch [182]) can produce continuous PPG data without any intervention from the user. PPG sensors take optical measurements from the skin surface and sub-surface to capture synchronous blood volume changes in the micro-vascular bed of tissues [183]. The signal PPG sensors produce thus is based on pulse transit dynamics at the wrist. This signal can be used to infer features of the pulse train, such as heart rate and heart rate variability (HRV). However, owing to its placement on the wrist and different signal dynamics, any inference from wrist-worn sensor data must navigate the challenges emanating from the differences between the two domains (chest vs. wrist). We briefly describe the three broad challenges with regards to adapting continuous inference models from the chest to wrist sensor domain. These challenges highlight the necessity of a distinct inference mechanism when using wrist-based sensor data.

Adapting Inference Models from Chest to Wrist

There are three essential distinctions between wrist-worn sensors and chest-based ones. The first is the wrists' location in the human body's periphery. Thus, sensing cardiac parameters becomes challenging owing to distance, and we have to make indirect measurements from the wrist. For example, chest-based ECG sensors directly assess the electric potential generated in the heart's lower ventricles. In contrast, wrist-based PPG sensors make heartbeat assessments from changes in blood volume in the wrist due to heartbeats. The second distinction concerns the degree of motion.

Compared to chest sensors, wrist sensors are susceptible to the significant movement of wrists which is dynamic and often unpredictable. Finally, a vital component of the sensor data collection concerns their attachment to the point of contact. Chest sensors usually employ sticky electrodes or conductive fabric-based chest-belt to attach themselves. Wrist sensors, on the other hand, are worn like watches. The firmness and placement of the watch-type sensor vary a lot from person to person, contributing to substantial changes in data quality. In the natural environment, dynamic wrist motion compounds these attachment concerns, thus restricting our ability to collect clean and reliable sensor data from wrist sensors. These challenges contribute to dynamic signal quality fluctuations of wrist-sensor data. The changes in signal quality significantly impact the robustness of inferences further down the line. Thus, quantifying the collected data's reliability is necessary before hypothesizing any subsequent inference mechanism on top of them. In translating an inference model from chest to wrist, we must consider two key elements: the ability of the sensing medium to collect reliable data for further inference and the accuracy of the inference itself.

This chapter aims to develop methods for robust inference of stress and activity using wrist-worn PPG and Accelerometry data. We first describe the activity recognition model we developed using labeled accelerometer data. We borrow from established literature on human activity recognition and develop a deep neural network model for muti-class activity recognition. Next, we focus our discussion on developing methodologies for continuous stress assessment from wrist-based PPG. We first establish the necessity of signal quality assessment of wrist-sensor collected PPG data. Next, we create a supervised learning-based data quality metric and integrate it within the stress assessment module. This approach makes it possible to satisfy the two critical requirements of continuous inference from wrist-sensor collected data. First, we assess the usability of the data to make reliable stress inferences by quantifying the signal

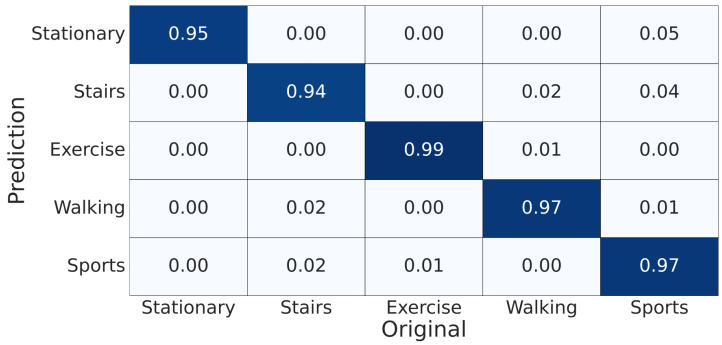


Fig. 4.1: Confusion Matrix of Activity Classification in WISDM dataset

quality level. Second, we improve the accuracy of stress inference itself by integrating our developed signal quality metric within the stress assessment methodology.

4.2 Activity Recognition From Wrist-worn Accelerometry

Wrist-fitted accelerometer sensors continuously capture the motion of the wrists in three orthogonal directions. We use inertial motion sensor data from wristwatches to compute participants' activity levels. Since activity is directly related to this motion, as presented in the IMU data, its inference is not highly susceptible to signal quality fluctuations and the quality of activity inference. We use a deep neural network based human activity recognition model. Deep learning models have become very popular for activity recognition due to their ability to encode and represent noisy sensor data to classify complex tasks [73]. Convolutional Neural Networks (CNN) offers the most efficient deep model architecture for activity classification in the wild [74, 75].

We train a CNN based activity recognition model for each 20-second data segment using publicly available WISDM dataset [76]. In WISDM, 51 participants performed 18 different activities while wearing accelerometers on their dominant wrists. Based on the amount of periodicity and variations present in different activity labels, we merge similar activities to obtain the following classes — *Stationary*, *Walking*, *Stairs*, *Sports*, and *Exercise*. *Stationary* refers to segments where the variation is minimum and encompass labels such as sitting, standing, typing and others. *Walking* incorporates

activities when there is gait information present, with those involving *Stairs* separated out. *Sports* refers to activities which consist of a mixture of stationary and sudden burst of active segments. These include playing, catching, dribbling, etc. *Exercise* includes activities of high magnitude such as jogging, running and cycling. Although periodicity is observed in the data segments for both *Exercise* and *Walking*, the two are different based on the magnitude of variations present.

For generalizing across orientation differences in different devices and study setups, we train the model using only magnitude of accelerometer data. Using 20% of each participants data as testing set, our model achieves an accuracy of 0.96 and a weighted F1-score of 0.96. Figure 4.1 shows the confusion matrix.

4.3 Robust Stress Inference from Wrist-worn PPG

Inferring stress from PPG depends on accurately assessing cardiac, and heart rate variability (HRV) features from wrist-worn PPG signals. However, due to their peripheral placement, dynamic wrist motion, and irregular attachment of wrist-worn sensors to the point of contact, PPG sensing in the natural environment suffers from various external noises and confounds. These noise elements become ingrained within the signals, diminishing their ability to reliably estimate the necessary cardiac and HRV features. It negatively impacts the quality of any subsequent inferences on top of PPG data. Therefore, robust stress inference from noisy wrist-worn PPG in the natural environment first requires decision-making on the state of the input PPG signal to contain valid information about participants' heart rate dynamics. In this section, we first establish the need to assess wrist-worn PPG data quality. We elaborate on the sources of noise present for wrist-based PPG data and the resultant challenges that manifest in PPG signal processing. We then describe the accuracy-yield tradeoff principle that applies to the continuous stress inference from wrist-worn PPG data in the natural environment. Next, we present the ideas behind the deep integration of signal quality metrics inside the stress inference module. Our methods leverage the benefits of continuous PPG

signal quality assessment in the natural environment and develop a robust stress inference pipeline from noisier wrist-worn PPG data.

Due to the optical nature of PPG sensing, PPG signal quality strongly depends on the sensor's attachment. In particular, the sensor's motion in the direction orthogonal to the skin surface can generate noise in the measured PPG waveform with amplitudes that can be as large as actual pulse transits. Unfortunately, periodic wrist motion due to walking and other forms of physical activity can result in significant noise in the time domain and strong noise components in the frequency domain that overlap with the typical range of valid heart rates. This noise can make analyzing PPG data more challenging as PPG signal quality can vary significantly over a day due to periods of physical activity and short periods due to transient hand motions that occur during other activities of daily living.

Notably, the difficulty in overcoming these challenges depends on the monitoring task of interest. In particular, different monitoring applications will have different sensitivities to the noise level in PPG data. For example, inference for heart rate from raw PPG data is relatively insensitive to noise. The dominant frequency in the PPG waveform over a given time window (e.g., one minute) will yield a reasonably robust estimate of heart rate unless there are significant noise components in the frequency domain due to periodic motion. Corruption due to motion at the wrist can be dealt with by leveraging actigraphy data collected at the wrist to determine when motion levels are low enough that PPG signal quality is likely to be in an acceptable range. This leads to a natural tradeoff between monitoring yield (defined as the number of minutes in which the system can produce usable inferences) and the accuracy of the monitoring output. Completely ignoring data quality considerations will generally result in the maximum yield and the lowest accuracy output. Restricting a system to produce results only when there is no motion will generally result in the lowest yield and the highest accuracy

output. The yield-accuracy tradeoff principle applies regardless of the monitoring task, as it is inherent to the PPG sensing modality under real-world deployment conditions.

To ensure reliable stress inference from wrist-worn PPG data facing the challenges mentioned herein, we emphasize the importance of developing an accurate and efficient PPG data quality metric that quantifies the amount of noise in the PPG signal irrespective of the sources. Since noise in PPG signals can often be transient, the ability to accurately and efficiently identify the level of corruption lets us determine the times of varying reliability and increased data uncertainty.

Assessing the signal quality levels also allows us to apply relative weighting to different locations of the PPG signal without discarding them altogether. This weighting mechanism diminishes the impact of transient noise and improves the robustness of computed cardiac and heart rate-based features from PPG signals. As a result, the accuracy and robustness of down-the-line inferences from the computed features improve substantially. Therefore, we propose integrating our developed signal quality metric within the stress inference mechanism. We thoroughly evaluate our developed signal-quality aware stress model in both lab and field environments. Our results show significant improvement in the accuracy of stress inference from wrist-worn PPG data compared to existing approaches. We also outline the improved accuracy-yield tradeoff profile of our developed stress inference pipeline.

4.4 Our Approach & Key Contributions

In this chapter, we first propose a new PPG data quality indicator for short time windows developed using supervised learning, which we refer to as CQP (for *See Quality of PPG*). We then show how CQP can be extended to longer time intervals and more deeply integrated into subsequent inferences to improve their robustness. The CQP data quality indicator is learned via an auxiliary classification task where the inputs are features extracted from a five-second segment of PPG data to capture potential rapid variation in PPG data quality. The goal is to decide if each segment is of acceptable

quality. A probabilistic classifier is used, and as a result, the CQP data quality indicator carries an interpretation corresponding to the probability that a segment of data is of acceptable quality. Using 28,000+ labeled PPG segments collected in the field, we show that CQP detects segments with acceptable data quality with 95% balanced accuracy compared to 80% using previous data quality measures. To assess the quality of longer segments (e.g., one minute) while being robust to incomplete data, we investigate different methods for aggregating the base CQP indicator over time.

To demonstrate the utility of CQP in complex inference tasks, we conduct a detailed case study comparing a PPG-based stress detection pipeline to an ECG-based pipeline. This study leverages unique paired ECG and PPG data from a lab study ($n = 36$ participants) and a field study ($n = 105$ participants). As in prior work using ECG, the primary PPG features of interest for this task are related to heart rate variability (HRV). As we will show, tasks that require more detailed PPG data analysis, such as the extraction of HRV feature from inter-beat time series, will typically exhibit a worse yield-accuracy profile than simpler statistics such as heart rate. However, our results indicate that higher PPG quality levels as given by CQP result in more accurate inference for all features investigated in this work. We then show how the CQP data quality indicator can be more deeply integrated into the PPG stress inference pipeline using a combination of minimum quality thresholding and quality-weighted feature computations, resulting in significantly improved accuracy-yield trade-offs.

We begin with a discussion of related work. We then describe the unique data sets we use in this work. Next, we turn to the development of the CQP model and its evaluation in the stress detection case study. We conclude with a discussion of limitation and future directions.

4.5 Related Work on PPG Signal Quality Assessment in Light of Physiological Event Inference

In this section, we discuss related work on PPG data quality assessment and the use of PPG data in physiological event inference.

4.5.1 Quality assessment of PPG Data

The majority of prior work on data quality assessment for PPG data applies to PPG sensors worn as a clip on the finger in clinical applications. Early work on PPG signal quality assessment developed thresholds on PPG morphological features to assess signal quality [184, 185]. PPG pulses were detected using repeated Gaussian filtering before applying thresholds on cross-correlation between consecutive pulse segments in [186]. Adaptive thresholding methods were developed to make quality assessments more robust. In [187], researchers used a beat-by-beat annotation of 'good' and 'bad' PPG pulse peaks to devise an online approach to classifying individual beats using a thresholding approach. In [95], authors use morphological characteristics with temporal variability information in the signal time series to assess the signal quality. However, classification based on thresholds developed on raw PPG signal still suffers from false beats, missing data, and significant beat variations in the field. Also, these works focus on binary assessment of acceptable PPG data, which leads to significant reduction of data volume in real life field conditions.

A number of approaches leverage template matching methods. For example, dynamic time warping (DTW) has been used to align each pulse beat to a running template for extracting signal quality indices in [188, 189, 110, 95]. By contrast, we focus on computationally simpler classification models applied to extracted features.

More recent work [190] investigated the effect of using a combination of several signal quality indices: perfusion [191], kurtosis [192], skewness [193], relative power, zero crossings, entropy, and the matching of systolic wave detectors. They manually annotated 106 PPG recordings each 60 seconds long into three distinct classes:

excellent, acceptable, and unfit. A 'Skewness index' was reported as the optimal quality feature, outperforming the perfusion index and other indices. 'Skewness' was further used by [194] as the single signal quality metric for designing an optimal filter for PPG signals.

Thresholds on motion derived from inertial sensors have also been used as a binary indicator for accepting PPG segments for further analysis [195, 179]. Considering motion as one of the indicators of signal quality, [196] proposed an accelerometer-based signal quality index. Similarly, [197] uses near-wrist and far-wrist based motion classification to remove segments of PPG affected by motion. Another work [198] uses accelerometer-based motion detection to ignore PPG segments affected by motion. We show that single indicators such as skewness or motion alone are insufficient to accurately identify acceptable segments of PPG data.

Several supervised machine learning approaches to signal quality assessment have also recently been proposed [190, 199, 200, 201, 202]. [190, 200] classify 60 second windows, while [199] classify 30 seconds. As window size increases, these methods miss the dynamic instantaneous fluctuations of signal quality. To be able to capture variations of signal quality as much as possible, we use 5 second windows to annotate signal quality. We also show the effect of aggregating assessed signal quality into longer windows. Further, most past work is based on a limited set of PPG data collected in controlled settings. This limits the sources of noise and opportunity to capture variations in signal quality deterioration present in real life conditions in the field. We annotated 28,086 PPG segments from 12 participants in the field setting and use these data to learn the proposed CQP PPG data quality indicator.

4.5.2 Using Signal Quality to Restore or Repair PPG Data

Much prior work has also dealt with the problem of corruption of PPG signals owing to hand or wrist motion. They focus on removing PPG segments affected by motion artefacts using conventional or adaptive filtering techniques [107, 108, 109],

template matching [110], wavelet transformation [111, 112], independent component analysis [113] and empirical mode decomposition [114].

Much of the existing work on PPG signal restoration is based on motion data collected from finger-based PPG sensors in bedside vital sign monitoring applications where motion is usually limited compared to the natural field environment. We incorporate the established knowledge of motion induced corruption in PPG signal when estimating heart rate information in the frequency domain using [82]. However, our proposed approach is also indicative of corruption not only due to motion but other factors prevalent in real life field conditions such as loose attachment, ambient light, power-line interference, and others. In contrast to the limited and constrained settings employed by most existing work, we collect data in both lab and the field setting. Our results show that in the field setting, inferences from PPG data can decrease sharply in accuracy as a function of quality compared to the lab setting.

4.5.3 Signal Quality for Physiological Event Inference

Prior work on inference for physiological events also often uses thresholds on signal quality to select data further analysis. For example, [203] used frequency thresholds to remove PPG segments affected by motion and noise and only use unaffected segments to identify coronary artery disease. Similarly, [204] used adaptive thresholds to remove noise and motion-affected PPG segments during a pre-processing step for estimation of mental distress from PPG. Work on atrial fibrillation detection [196, 205, 206] also uses thresholded indicators of signal quality to remove low-quality segments. Past work on stress assessment, drug use detection, and pain detection from PPG similarly excludes noisy segments of PPG data [103, 88, 179]. We propose a method that uses both quality weighting and quality thresholding to produce an improved yield-accuracy trade-off compared to only using quality thresholding as investigated in these past approaches.

(a) Lab study protocol (b) Chest Band (c) Wrist Band

Fig. 4.2: Lab study protocol and devices used for data collection. The chestband consists of ECG, respiration, and accelerometers. The wrist device consists of 3 channels of PPG, accelerometers, and gyroscope.

4.6 Datasets

Our goal in this chapter is to learn a data quality indicator for wrist-worn PPG and evaluate its accuracy and utility when integrated into complex real-world inference pipelines. We compare to established approaches based on ECG data. To enable this work, we leverage data sets collected via unique studies that include paired ECG and PPG data collected in both the lab and the field settings. As noted in the introduction, we use stress detection as a case study. Thus, the lab study specifically collects data under a controlled stress inducement protocol. This provides unambiguous stress labels for training and testing stress detection models. The field study provides ecologically valid data from PPG and paired ECG. We compare the output of PPG-based models to those of ECG-based models in the field to assess the performance of PPG-based models. Both studies were approved by the local Institutional Review Board and all participants provided written consent. We now describe details of the devices, study protocol, and data collected.

4.6.1 Devices

In both studies, participants wore a chest-band device that included ECG, respiration, and accelerometer sensors. Participants in both studies simultaneously wore identical wristband devices on both wrists that included a PPG sensor and 6-axis inertial sensors. Data was collected and transmitted to study servers via a smartphone. We describe the devices below.

- **Chest-band Device:** The device consits of a flexible chest belt (see Figure 4.2b) with a two-electrode ECG sensor, respiratory inductive plethysmography (RIP) sensor for measurement of relative lung volume, and a 3-axis accelerometer to assess the motion of the torso. The sampling frequency of sensors is 100 Hz for ECG and 25 Hz for both respiration and accelerometer in the lab. In the field, the sampling frequencies are 64 Hz for ECG, 21.33 Hz for both respiration and accelerometer.

- **Wristband Device:** In both lab and field, the collection of PPG was performed using two wristband (one worn on each wrist), as shown in Figure 4.2c. It captures PPG signals in 3 different LED channels (red, infrared and green) using two receivers along with a 3-axis accelerometer and gyroscope. The sampling frequency is 25 Hz for all the sensors in all settings.

- **Smartphone:** The wristband and the chestband sensors transmit the collected sensor data wirelessly to an Android smartphone in real-time via Bluetooth Low Energy (BLE). The smartphone timestamps all received sensor data. Timestamp correction of received bytes occurs in the smartphone software [156]. The smartphone software uploads data to the cloud wherever bandwidth is available. Participants in the field study were provided with a study phone configured with all the necessary software.

4.6.2 Study Protocols

The protocol for the lab study was designed to replicate stress situations and was conducted in a controlled environment. The field study was designed to investigate the role of stress and environmental cues (detected by sensors) in triggering smoking lapse in newly abstinent African American smokers [30, 207].

Lab Study Protocol

The lab study protocol is based on [208], which showed that cardiovascular and neuro-endocrine adjustment to public speaking and mental arithmetic exhibit stress response in physiology. The protocol replicates that of [15]. The study was designed to subject participants to three types of validated stressors — socio-evaluative (public speaking preparation & delivery), cognitive (mental arithmetic), and physical (dipping hands in ice cold water) in a repeated measure design. The study consisted of a 30 minute baseline period where participants were asked to sit and rest. The socio-evaluative challenge consisted of a preparation phase (4 minutes) and a speech delivery phase (8 minutes). The cognitive challenge components consisted of a mental arithmetic session of increasing difficulty (4 minutes). Finally, for the physical challenge, participants were asked to submerge their hand in ice cold water for 90 seconds. Lab study sessions ended with a 30 minute stress recovery session. Participants were given instructions before each session. Figure 4.2a shows the sequential lab study session design. The tasks depicted in red have been shown to induce stress-related physiological changes [208], whereas tasks in green represent rest or recovery time periods. Each distinct rest and stress period was timestamped to create ground-truth labels for each minute of the session similar to [15]. Instruction periods between the consecutive tasks are not taken into consideration for ground-truth labels.

Field study Protocol

The participants for the field study were recruited for a smoking cessation research project. To be eligible, participants had to be smokers for two years, have no ongoing medical or psychiatric illness, and have a willingness to quit. After being enrolled at the baseline visit, participants were trained in the proper use of the sensor devices and how to respond to questionnaires in the form of Ecological Momentary Assessments (EMA) via mobile phones. The participants wore both the chest and wrist

Table 4.1: Description of data from the Lab & Field Studies

	Participant Information			Data Volume (hours)					
				Chest (ECG)		Wrist (PPG, ACL)			
						Left		Right	
Study	No. of Participants	Participants Selected	Age	Total	Acceptable	Total	Not Irrecoverable	Total	Not Irrecoverable
Lab	39	36	—	122	39	66	43	66	43
Field	131	105	52.19±11.43	18,850	4,706	16,245	10,430	15,879	10,069

sensor suites during waking hours. Each participant was enrolled in the study for 14 days. The participants were compensated for their time and effort.

4.6.3 Data Collected

Table 4.1 summarizes the data collected in both studies. Participants for the lab study were recruited via printed advertisement (e.g., flyers, local weekly periodicals), online advertisement using social media and list groups, announcement in classrooms, and word of mouth. A total of 39 unique participants completed the lab study consisting of a single session. Three of the 39 participants had no usable data and hence were excluded from further analysis. From the remaining 36 participants in the lab, we have 132 hours of PPG data belonging to lab protocol sessions from both wrists combined (66 hours from right wrist, 66 hours from right wrist). After excluding irrecoverable segments (segments from which heart rate information can not be recovered, see Section 4.7.3 for more details), we retain 43 hours of acceptable PPG data from left wrist and 44 hours from right wrist. We get 39 hours of acceptable ECG data.

The field study recruited participants via print advertisement (e.g., flyers, local weekly periodicals) and advertisement on radio. The phase of the study with paired ECG and PPG data enrolled 131 unique participants, of which 21 participants withdrew from the study. Out of the remaining 110 participants, we have paired ECG & PPG data from 105. The mean age of the participants is 52.19 years with a standard deviation of 11.43 years. From these participants, we have 16,245 hours of PPG data collected from the left wrist and 15,879 hours of PPG data from the right wrist. After filtering out irrecoverable segments and only including the days when we have at least 1 hour of recoverable data present (necessary for stress inference that needs baseline data for

day-specific normalization), we have data from 1,095 unique participant days belonging to 105 participants. Total recoverable PPG data amounts to 10,430 hours from the left wrist and 10,069 hours from the right wrist. We have 18,850 hours of ECG data in the field study, out of which chest sensor was worn on the body for 7,604 hours. Amongst the on-body hours, we have acceptable ECG data for a total of 4,706 hours. Around 3,000 hours of ECG data is unusable due to intermittent packet loss and loose attachment of electrodes. We use these data to learn and evaluate the proposed PPG data quality index.

4.7 CQP Data Quality Model

We first describe the data pre-processing, feature extraction, data labeling, learning and evaluation of the proposed CQP model for inferring PPG data quality over five-second windows, which we refer to as CQP-5. We then turn to the question of leveraging CQP-5 to define a PPG data quality indicator over 60 second windows, CQP-60.

4.7.1 Windowing Data for Quality Assessment

The first step in the assessment of signal quality is to select the size of window to which data quality assessments should be ascribed. Selecting a longer window size will increase the quality variation within the window, whereas smaller window sizes may not contain enough data to assess signal quality reliably. We use the recommendation from [194] that suggests a minimum of 5 seconds as the necessary window size to accurately discriminate acceptable segments from unacceptable segments. Five second windows also gives us sufficient frequency and time resolution to calculate heart rate information within individual windows. We thus begin by developing the base CQP model, CQP-5 for five second windows of PPG data. We subsequently turn to the problem of how to aggregate the output of CQP-5 over longer windows and develop CQP-60, which provides a data quality index for one minute (60 second) segments of

data. CQP-5 and CQP-60 are both used in the stress inference task. We next discuss data pre-processing steps for the CQP-5 model.

4.7.2 Preprocessing PPG Data

Before assessment of signal quality, we clean the PPG data by filtering out high-frequency noise. We employ a 64^{th} order Finite Impulse Response Butterworth bandpass filter for removing high-frequency noise. This filter retains frequencies ranging from 0.4 Hz to 3.5 Hz (24 - 210 BPM). Using bandpass filtering allows us to filter out contributions from external noise sources which are prevalent in unwanted frequencies while not affecting the information content from the pulse variations associated with heart activity.

PPG signals collected under uncontrolled field conditions can also have large and rapid variability in amplitude due to varying proximity to skin and the resulting response of the gain controller included in the sensor hardware. Therefore, to make amplitude variations more consistent, we normalize the PPG data using quartile-based normalization. We normalize each 5 second PPG segment by centering it using the median value over the interval and scaling it using the inter quartile range. If $X = [x_1, x_2, \ldots x_k]$ denotes a PPG segment of length k in a particular channel, then we normalize X using the transformation shown below:

$$X_{\text{norm}} = \frac{X - Q_2(X)}{Q_3(X) - Q_1(X)}, \tag{4.1}$$

where Q_n denotes the n^{th} quartile. Normalization reduces the effects of outliers within a segment and standardizes the computed features. The inter quartile range and the median as measures of spread and location are selected instead of the more common Z-transform as they are more robust to outliers.

4.7.3 Identifying and Isolating Irrecoverable PPG Segments

As with many applications of wearable sensors, it is possible for the data expected in PPG signals to be missing or so corrupted by noise that the data are not usable. We refer to segments of time where this is the case as being *irrecoverable*. These may correspond to situations when the sensor is not worn leading to missing data, where there is high momentary noise, or when the sensor is worn too loosely to capture pulse peaks. Therefore, as our next data processing step, we detect and remove irrecoverable windows of data.

We operationalize the concept of irrecoverable windows of data via spectral analysis. Specifically, if there are no power spectral peaks at all within the heart-rate frequency range (0.8 - 2.5Hz), we deem the segment to be irrecoverable. The motivation for this definition stems from that fact that when it is violated, standard spectral methods for heart rate estimation would fail to identify a heart rate value for the segment [83].

To compute the condition, let *S(f)* be the normalized power spectrum calculated using the Welch's method [209]. We let *I(f_p)* indicate whether f_p is a peak in the power spectrum as defined below. We consider a PPG segment to be irrecoverable if $\sum_{f=f_{min}}^{f_{max}} I(f) = 0$.

$$I(f_p) = \begin{cases} 1, & if \ \frac{d}{df}\Big|_{f=f_p} S(f) = 0 \quad and \quad S(f_p) \geq c \cdot \max_f S(f) \\ 0, & otherwise \end{cases} \tag{4.2}$$

We select the threshold *c* empirically from our data. In [210], the authors considered any potential peak in power spectral density as a spectral peak if its amplitude was at least 30 percent of the maximum. Our goal is for this stage to have a low false negative rate and a high true positive rate for identifying segments that are recoverable. We select c = 0.1 for our experiments, which provides sufficient leeway to accept noisier

segments that may still be usable. The signal quality model is applied to all segments that are not deemed to be irrecoverable and can be used to provide a second layer of more fine-grained thresholding. We next turn to a discussion of feature extraction for the quality assessment model.

4.7.4 PPG Signal Quality Features

Several features have been proposed to assess the quality of PPG signals in prior research [190, 201, 199]. They can be classified into two types: those based on time-domain representation and those based on frequency domain representations. Time-domain features include perfusion index, skewness, kurtosis, zero crossings, standard deviation, mean, median, and inter-quartile range. These features mostly define the shape and symmetry of the distribution of PPG values within a time window. Perfusion index is defined by the relative range of bandpass filtered PPG compared to raw value and is used in commercial smart-watches [211]. In terms of frequency domain metrics, we focus on relative power defined as the ratio of the power spectral density (PSD) in the heart rate frequency band compared to the PSD in the overall signal. In [190] skewness was found to be the the optimal signal quality index for classification of PPG signals collected from finger-based PPG sensors in the lab environment.

We select four features for our model that are not easily affected by potential sources of between-person variability (such as skin color). They are skewness, kurtosis, relative power, and standard deviation of normalized PPG segments. Skewness measures the symmetry in the distribution of PPG data within a window. Kurtosis measures the probability in the tail of PPG data distribution within a window. Standard deviation of PPG segments normalized by quantiles according to (4.1) is used as a robust metric to measure per sample deviation. We use relative power to measure the contribution of frequencies belonging to the heart rate range relative to those outside that range. Relative power is defined in [190] as $R_{\text{power}} = \int_{f_{min}}^{f_{max}} S(f)df / \int_{0}^{\infty} S(f)df$, where $S(f)$ is the normalized power spectrum. We select f_{min} = 0.8Hz and f_{max} = 2.5Hz corresponding

to 48 BPM and 150 BPM respectively, the frequency band where we expect heart rates to be located under normal conditions.

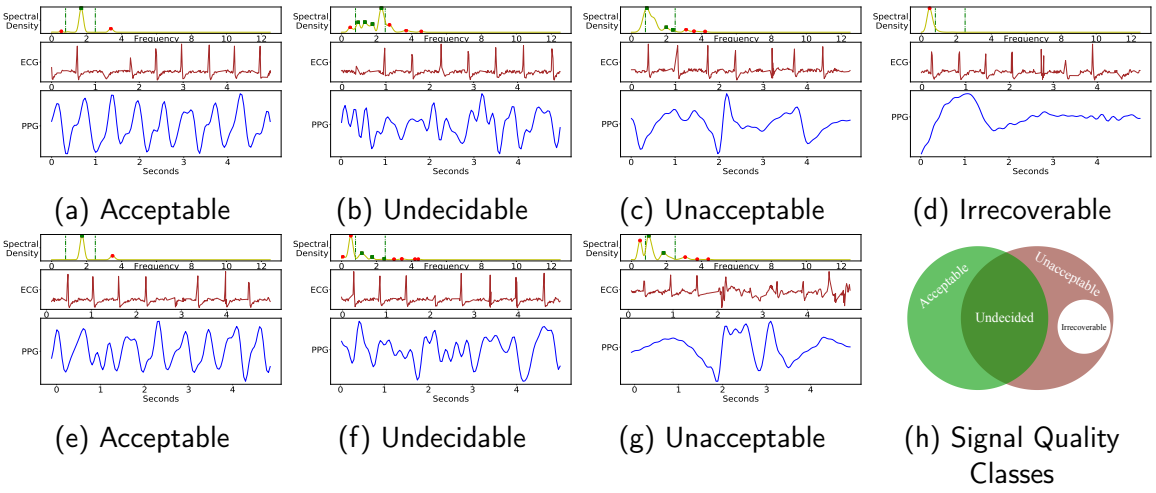
We also consider inertial motion features used in studies employing commercial watches, which cease sampling of PPG signals or discard it from further analysis when sufficient motion is present [105, 195]. We evaluate this approach compared to the proposed approach and also consider using these features in the proposed model. In the next section, we turn to the question of annotating PPG data segments with data quality labels for use in data quality model learning.

4.7.5 PPG Data Quality Labeling

To train and test a classifier for assessing the quality of PPG data collected in the natural field environment, we take a similar approach to [194]. We annotate five-second windows of field data with labels from the set {Acceptable, Undecidable and Unacceptable} for subsequent analysis. We define these three classes below. Figure 4.3h shows the relationship among the classes (including the irrecoverable class defined earlier) in a Venn diagram.

- **Unacceptable for analysis**: We consider a PPG segment *unacceptable* if no prominent systolic peak is present in the time domain as well as the absence of PPG pulses with appropriate morphology and corresponding lag relationship with the ECG R peaks. There can be scenarios where spectral peaks are present in the heart rate frequency range. But, their contribution to time-domain periodicity is quite limited.

- **Acceptable for analysis**: If a PPG segment is not unacceptable, we look into the time-domain representation of the segment. If there is a dominant beat morphology containing systolic peaks, which corresponds to at least two full PPG pulses of duration between 400 ms and 1,250 ms and systolic peaks show correspondence with ECG R peaks, we consider the segment as *acceptable*.



(a) Acceptable (b) Undecidable (c) Unacceptable (d) Irrecoverable

(e) Acceptable (f) Undecidable (g) Unacceptable (h) Signal Quality Classes

Fig. 4.3: Figures a-g show annotation of single channel PPG windows with respective power spectral density plot on top. The vertical lines in the top subplot show the heart rate frequency range. Spectral peaks inside and outside this range are colored separately. Figures (a) and (e) show acceptable segments, (b) and (f) show undecidable segments, (c) and (g) show unacceptable segments and (d) shows irrecoverable segments Figure (h) shows the intersection of different classes assigned to windows through a Venn Diagram. Irrecoverable subclass is shown inside a gap within the Unacceptable class.

- **Undecidable**: If PPG segment cannot be categorized based on the classes as mentioned above, we label it as *undecidable*. Annotation into undecidable class gives space for the ambiguity generated by the presence of quasi-periodicity in the time domain. A non-prominent spectral peak in the frequency domain can represent periodicity in the time domain. However, the interpretation of the time-domain peak as acceptable warrants the presence of a PPG pulse of the required duration (400ms to 1,250ms). Also, to confirm the presence of a PPG pulse, we need a prominent systolic peak that can be attributed to the pulse. The systolic peaks also need to be in sufficient agreement with ECG R peaks. In such cases, it may not be clear if the PPG data are valid or not.

The annotation was performed by two independent annotators. The annotators based their label decisions on visualizations of a normalized time-domain representation and a power spectral density representation of the raw PPG data, since these representations are closer to the feature values that a model will subsequently access

Table 4.2: Inter-rater distribution statistics for calculation of kappa statistic κ

	Acceptable	Undecidable	Unacceptable
Acceptable	6202	1127	371
Undecidable	1422	5570	1662
Unacceptable	269	1548	9915

than the raw PPG data. To aid in the labeling process, each segment of PPG data also has a corresponding segment of ECG data displayed. The inclusion of ECG facilitates the annotation of PPG segments and provides clean reference information to the annotators. Example visualizations for each class are shown in Figure 4.3.

After discarding all the irrecoverable PPG segments, we randomly select 28,086 single-channel 5-second PPG segments (for a total of 2,340 minutes, or 39 hours) from 12 participants in the field, each of length 5 seconds. We normalize each window according to (4.1). Two independent annotators classified each instance according to the guidelines mentioned above. We compute the inter-rater agreement between the two annotators using Cohen's kappa value. The unweighted kappa coefficient calculated using [212] is 0.652 with a 95% confidence interval range of (0.645, 0.659). This is considered a substantial agreement. Also, this is significantly better than the value of 0.48 reported in [190] based on 106 PPG segments. We note that the inclusion of an Undecidable class gives us some leeway when faced with an ambiguity of labeling windows into binary classes of acceptable and unacceptable. We next describe the CQP-5 model learned using these labeled data.

4.7.6 The CQP-5 Model for Assessment of PPG Data Quality Over Five-Second Windows

The goal of the CQP-5 model is to provide a continuous signal quality index such that higher values of the index correspond to data of that are increasingly likely to be of acceptable quality. We frame this as a binary classification problem where Acceptable is the positive class and Unacceptable and Undecidable are collapsed together to form the

negative class. The probability of the positive class then corresponds to the required signal quality index.

The CQP-5 model is applied to five-second segments of data. We consider all of the features described in Section 4.7.4 including Skewness, Kurtosis, Relative Power, and Standard Deviation computed from the PPG data, as well as standard deviation of accelerometer magnitude computed from the associated actigraphy data. We consider learning the CQP-5 model using two different model classes: logistic regression with ℓ_2 weight decay (LR-CQP-5) and decision trees (DT-CQP-5). For logistic regression we optimize the regularization hyper-parameter. For the decision tree model class, we optimize the criteria to split a node in a decision tree. To account for the imbalance in classification we optimize the class weight parameter for both logistic regression and decision tree. We perform hyper-parameter optimization using a grid-search cross validation approach.

We select the logistic regression and decision tree model classes as computationally efficient linear and non-linear models instead of more sophisticated model classes as we are seeking a computationally-efficient classifier that can be easily deployed in the field for frequent real-time assessment of PPG signal quality with limited resources. In the next section, we look at the performance of both models.

4.7.7 Five-Second PPG Data Quality Classification Experiments

In this section, we evaluate the performance of the proposed CQP-5 model compared to a range of simpler threshold models computed from individual features. As noted in the previous section, we use Acceptable as the positive class and group together Unacceptable and Undecidable as the negative class. Further, any instances where there was disagreement between the annotators were placed in the negative class. This resulted in 1,978 instances of the positive class and 11,893 instances of the negative class. We use a 66/34 train-test split and apply 10-fold stratified cross-validation within the training set to optimize hyper-parameters.

The specific models that we consider are listed below. The threshold classifier (TC) models learn a lower and upper threshold on a single feature to discriminate the positive class from the negative class. For any feature with value f, we find two thresholds α and β such that when $\alpha \leq f \leq \beta$ the output is positive and otherwise the output is negative. We find the values of α and β for each feature by optimizing the balanced accuracy using 10 - fold stratified cross validation applied to the training data.

- TC-Sk-5: A threshold classifier model based on skewness of the PPG data.
- TC-Kr-5: A threshold classifier model based on kurtosis of the PPG data.
- TC-RP-5: A threshold classifier model based on relative power of the PPG data.
- TC-SDP-5: A threshold classifier model using the standard deviation of the PPG data.
- TC-SDA-5: A threshold classifier model using the standard deviation of accelerometer magnitude.
- DT-SDA-5: A decision tree model using the standard deviation of accelerometer magnitude.
- LR-CQP-5: The CQP-5 model based on the logistic regression classifier.
- DT-CQP-5: The CQP-5 model based on the decision tree classifier.

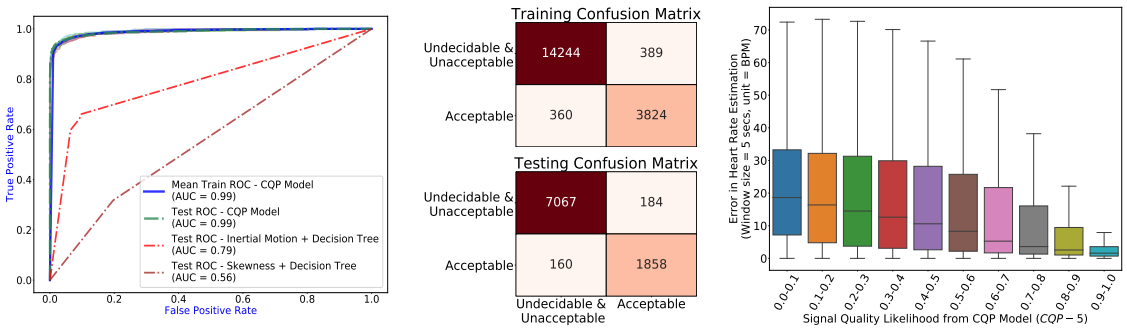
Results comparing these models are shown in Table 4.3. We find that the model based on Skewness only (TC-Sk-5) has the lowest F1 and balanced test accuracy. Interestingly, this feature was found to be optimal when detecting the quality of a finger-clip based PPG sensor in a controlled environment [190]. Of the four single-feature PPG -based models tested, TC-RP-5 based on realtive power achieves the best test F1 score at 0.8.

Table 4.3: PPG Signal Quality Model Performance on 5-second Segments

Model	10-Fold Stratified Cross Validation Training Results				Testing Results			
	F1 Score	Balanced Accuracy	Precision	Recall	F1 Score	Balanced Accuracy	Precision	Recall
TC-Sk-5	0.5	0.71	0.35	0.89	0.5	0.71	0.35	0.89
TC-Kr-5	0.74	0.85	0.7	0.79	0.74	0.84	0.69	0.78
TC-RP-5	0.8	0.88	0.76	0.84	0.8	0.87	0.75	0.85
TC-SDP-5	0.6	0.6	0.42	0.98	0.6	0.6	0.42	0.98
TC-SDA-5	0.57	0.74	0.5	0.66	0.57	0.74	0.5	0.67
DT-SDA-5	0.64	0.77	0.64	0.64	0.68	0.8	0.66	0.69
LR-CQP-5	0.89	**0.95**	0.82	**0.96**	0.89	**0.95**	0.82	**0.96**
DT-CQP-5	**0.91**	0.94	**0.91**	0.91	**0.92**	**0.95**	**0.91**	0.93

Next, we evaluate the classification performance when using motion features alone. We can see that these models (TC-SDA-5 and DT-SDA-5) outperform the skewness feature, but are substantially worse than the PPG relative power model (TC-RP-5). The lower performance when using motion features alone may be because even though motion may reduce the quality of PPG signals, several other factors can decrease the quality of PPG signals such as loose attachment and ambient lighting.

Lastly, we turn to the proposed models: LR-CQP-5 and DT-CQP-5. We can see that both of these models, which leverage all features, significantly outperform the other models. Of the two models, the decision tree model obtains the best test F1 value at 0.92 and ties the logistic regression model for test balanced accuracy at 0.95. Due to the poor performance of the motion features when used individually, we performed an ablation study to determine the effect of removing the motion features. This resulted in the DT-CQP-5 obtaining exactly the same performance as when motion features are used. As a result, in the remainder of this paper, we use DT-CQP-5 computed using PPG features only to provide the 5-second data PPG quality index CQP-5. Before turning to the derivation of a CQP data quality indicator for longer segments, we first experiment with how CQP-5 performs on the task of stratifying heart rate inferences by data quality.



(a) (b) (c)

Fig. 4.4: Signal Quality Classification Results.Figure (a) shows train and test ROC for CQP model, test ROC for skewness and inertial motion using decision tree model. Figure (b) shows train and test confusion matrices for two classes: Undecidable/unacceptable, and acceptable. Training confusion matrix is obtained after 10 fold-stratified cross validation applied to training data. Figure (c) shows the error distribution in heart rate estimation in different bins of signal quality likelihood for each 5-second segment of PPG data.

4.7.8 Five-Second Instantaneous Heart Rate Estimation Experiments

In this experiment, we consider an application of the CQP-5 model to the problem of instantaneous heart rate computation. Starting with the IEEE signal processing cup in 2015 [78], there has been a surge of work on heart rate estimation from PPG. Several of these methods [83, 79, 213, 214, 77] use spectral analysis of PPG, assisted by inertial signal analysis. For the classical spectral peak detection, we adopt the model from [82]. Power spectral density of input PPG channel is calculated using Welch's method [209] and the spectral peak frequency (calculated using (4.2)) belonging to heart rate range (0.8Hz - 2.5Hz) with maximum amplitude, and not considered as motion artifact, is selected as the output heart rate frequency. The frequency axis is linearly related to heart rate in beats per minute, i.e., $HR = 60 * f$, where f is the dominant peak frequency in Hz. Mean RR interval in milliseconds is calculated using the formula $MeanRR_{int} = \frac{60000}{HR}$. Figure 4.4a shows the error distribution for PPG-based heart rate estimation in different CQP-5 data quality bins for each 5-second segment of PPG data with heart rate calculated from ECG beat to beat interval used as ground truth. Results are shown for 8,900,962 windows from 46

participants in the field. As expected, the median, maximum and spread in errors decreases monotonically as a function of the CQP-5 data quality indicator.

4.7.9 The CQP-60 Index for Assessment of PPG Data Quality Over Sixty-Second Windows

Inference for higher-level health states typically uses a larger window of data than the 5-second intervals the CQP-5 model is based on. For example, past work on stress inference uses one minute window of data [15]. In this section, we address the problem of computing a data quality index over longer time intervals. As described earlier, the CQP-5 model outputs the probability that a 5-second window of data is of acceptable quality. While we could apply the same methodology to define a model over longer windows of data, there can be sufficient variation in data quality within longer windows that this approach would lack flexibility for. Instead, we consider aggregating the output of the CQP-5 model through time to define the signal quality of longer windows. In this section, we focus specifically on 60-second windows and define the CQP-60 index.

To derive the CQP-60 index, we first compute the output probability of the CQP-5 model for 5 second windows within the larger 60-second segment. We slide the 5 second window by 2.5 seconds, resulting in 50% overlap between adjacent 5-second windows. This results in 23 overlapping 5-second windows within each minute. Each 5 second window contains three channels of data. CQP-5 is applied independently to each channel and we dynamically select the channel with the highest signal quality for subsequent computation. However, some 5-second windows of PPG data within the larger one minute window can also be unavailable for analysis due to being labeled as irrecoverable in earlier stages of processing. A one-minute window of PPG data is considered to have a quality of -1 if less than p percentage of expected windows are present. ECG-based stress models use $p = 66\%$ [15]. We instead select $p = 50\%$ since irrecoverable data occurs fairly frequently in our PPG field data.

For one-minute windows of data with at least $p = 50\%$ of data points available,

we compute the CQP-5 model's probability output over the available data and then aggregate the obtained values to form a quality index for the overall minute of data. We consider several possible aggregation functions including the minimum, median, and mean of the available CQP-5 outputs in the window. We refer to these index values as Mean-CQP-60, Median-CQP-60 and Min-CQP-60. In the next section, we evaluate these three options on a range of PPG feature computations relevant to stress detection to select an optimal aggregation function.

4.7.10 60-Second PPG Feature Computation Experiments

To select an appropriate aggregation function, we compute each of the CQP-60 index variants described in the previous section (Mean-CQP-60, Median-CQP-60 and Min-CQP-60) for each minute of PPG field data. We also use the corresponding PPG field data to compute a variety of useful PPG features of different complexity. We assess the correlation between the output of the PPG-based computation of these features and the ECG-based computation of the same features over the same windows using both lab and field data stratified by minimum data quality using each CQP-60 model variant. We deem one aggregation function to be better than another to the extent that its corresponding 60-second quality index provides a better yield-correlation trade-off.

The features we use in these experiments are useful for a variety of tasks, including inference for stress. For first order features, we compute the heart rate and 80^{th} and 20^{th} percentile of the mean RR interval timeseries in a minute. For second-order features, we compute the interquartile range (IQR) and high-frequency energy (0.3-0.4 Hz). We select these features because they are discriminative for stress inference [15]. As the signal quality threshold increases, the amount of data that meet the threshold decreases. Therefore, we also compute the data yield corresponding to each of these thresholds. To observe the lab to field generalizability, we compute the correlation and yield for both lab and field data.

Table 4.4 and Table 4.5 present the results of this experiment in lab and field

Table 4.4: Correlation of Minute level HRV features computed from PPG in Lab

Data Quality Index	Threshold	Lab					
		Feature Correlation					Yield
		Heart Rate	80th Percentile	20th Percentile	IQR	HF Energy	Percent Reduction in Yield
Mean-CQP-60	≥ 0.0	**0.69±0.22**	**0.68±0.23**	**0.71±0.21**	0.43±0.16	0.44±0.17	100.00±18.50
	≥ 0.1	**0.69±0.21**	**0.71±0.22**	**0.74±0.20**	0.45±0.17	0.44±0.16	91.40±19.32
	≥ 0.2	**0.79±0.20**	**0.77±0.21**	**0.76±0.21**	0.44±0.17	0.49±0.18	77.99±20.44
	≥ 0.3	**0.86±0.20**	**0.85±0.20**	**0.81±0.21**	0.52±0.20	0.51±0.18	66.88±20.17
	≥ 0.4	**0.88±0.14**	**0.87±0.16**	**0.84±0.19**	0.56±0.20	0.58±0.16	58.55±19.40
	≥ 0.5	**0.91±0.15**	**0.91±0.17**	**0.85±0.22**	0.57±0.23	0.63±0.17	50.55±18.64
	≥ 0.6	**0.93±0.15**	**0.92±0.19**	**0.85±0.18**	**0.66±0.22**	**0.69±0.18**	44.54±17.66
	≥ 0.7	**0.95±0.08**	**0.93±0.11**	**0.91±0.15**	**0.72±0.19**	**0.70±0.16**	38.28±16.17
	≥ 0.8	**0.96±0.12**	**0.93±0.19**	**0.95±0.12**	**0.72±0.20**	**0.74±0.20**	31.99±12.89
Median-CQP-60	≥ 0.0	**0.69±0.22**	**0.68±0.23**	**0.71±0.21**	0.43±0.16	0.44±0.17	100.00±18.50
	≥ 0.1	**0.81±0.20**	**0.83±0.21**	**0.76±0.21**	0.43±0.18	0.48±0.18	74.10±21.07
	≥ 0.2	**0.84±0.19**	**0.86±0.19**	**0.81±0.18**	0.49±0.19	0.54±0.16	64.74±20.25
	≥ 0.3	**0.88±0.16**	**0.88±0.17**	**0.84±0.19**	0.54±0.19	0.57±0.16	58.36±19.48
	≥ 0.4	**0.89±0.16**	**0.89±0.17**	**0.83±0.18**	0.56±0.20	0.59±0.16	54.83±18.81
	≥ 0.5	**0.91±0.15**	**0.90±0.18**	**0.86±0.21**	0.57±0.23	0.62±0.17	50.49±18.70
	≥ 0.6	**0.93±0.16**	**0.92±0.19**	**0.85±0.21**	**0.61±0.21**	**0.64±0.16**	47.03±18.49
	≥ 0.7	**0.92±0.15**	**0.91±0.19**	**0.85±0.19**	**0.61±0.21**	**0.64±0.15**	44.97±17.82
	≥ 0.8	**0.94±0.09**	**0.92±0.12**	**0.89±0.16**	**0.69±0.17**	**0.69±0.18**	40.73±17.41
Min-CQP-60	≥ 0.0	**0.69±0.22**	**0.68±0.23**	**0.71±0.21**	0.43±0.16	0.44±0.17	100.00±18.50
	≥ 0.05	**0.97±0.15**	**0.93±0.20**	**0.95±0.15**	**0.73±0.18**	**0.72±0.17**	27.22±10.95
	≥ 0.1	**0.97±0.22**	**0.94±0.21**	**0.95±0.23**	**0.72±0.23**	**0.74±0.22**	23.37±10.00
	≥ 0.15	**0.97±0.14**	**0.94±0.18**	**0.95±0.16**	**0.75±0.21**	**0.75±0.22**	21.89±8.19
	≥ 0.2	**0.96±0.13**	**0.95±0.19**	**0.96±0.11**	**0.80±0.18**	**0.78±0.22**	21.94±7.00

respectively. We make several interesting observations that can inform the selection of appropriate signal quality aggregate function as well as specific operating thresholds. First, using Mean-CQP-60 and Median-CQP-60 result in a significantly higher yield than Min-CQP-60. For example, the yield at quality level 0.2 for Min-CQP-60 is the same as that for quality level of 0.8 for Mean-CQP-60. This can be explained by the fact that minimum threshold is quite stringent as for any minutes to qualify, all five-second sub-windows must individually meet the same threshold. Second, we observe that among Mean-CQP-60 and Median-CQP-60, the performance is similar, with marginally better yield for Median-CQP-60. But, Mean-CQP-60 exposes a larger range of correlations and yield and this provides a greater range of useable operating points compared to Median-CQP-60. Therefore, we select Mean-CQP-60 for our subsequent analysis.

Table 4.5: Correlation of Minute level HRV features computed from PPG in Field

Data Quality Index	Threshold	Field						
		Feature Correlation					Yield	
		Heart Rate	80th Percentile	20th Percentile	IQR	HF Energy	Minutes Per Wrist Day	Percent Reduction in Yield
Mean-CQP-60	≥ 0.0	0.46±0.15	0.41±0.16	0.50±0.15	0.14±0.05	0.14±0.05	606.00±168.75	100.00±27.85
	≥ 0.1	0.52±0.15	0.45±0.16	0.54±0.14	0.14±0.05	0.16±0.05	519.80±158.19	85.78±26.10
	≥ 0.2	**0.61±0.15**	**0.60±0.17**	**0.62±0.14**	0.16±0.06	0.18±0.05	384.16±148.62	63.39±24.52
	≥ 0.3	**0.70±0.15**	**0.70±0.17**	**0.71±0.14**	0.20±0.08	0.21±0.06	283.16±133.70	46.73±22.06
	≥ 0.4	**0.76±0.16**	**0.72±0.18**	**0.76±0.15**	0.21±0.09	0.23±0.07	190.39±106.01	31.42±17.49
	≥ 0.5	**0.80±0.17**	**0.75±0.19**	**0.80±0.16**	0.24±0.11	0.28±0.09	141.28±82.30	23.31±13.58
	≥ 0.6	**0.85±0.16**	**0.79±0.19**	**0.83±0.16**	0.30±0.12	0.35±0.11	124.33±67.93	20.52±11.21
	≥ 0.7	**0.89±0.15**	**0.83±0.16**	**0.87±0.14**	0.33±0.14	0.33±0.14	112.88±59.50	18.63±9.82
	≥ 0.8	**0.92±0.13**	**0.84±0.15**	**0.91±0.12**	0.33±0.16	0.36±0.17	102.68±52.89	16.94±8.73
Median-CQP-60	≥ 0.0	0.46±0.15	0.41±0.16	0.50±0.15	0.14±0.05	0.04±0.02	606.00±168.75	100.00±27.85
	≥ 0.1	**0.60±0.15**	**0.58±0.17**	**0.61±0.14**	0.16±0.06	0.09±0.04	381.58±150.14	62.97±24.78
	≥ 0.2	**0.67±0.15**	**0.66±0.17**	**0.68±0.14**	0.21±0.08	0.12±0.05	311.76±143.16	51.45±23.62
	≥ 0.3	**0.71±0.16**	**0.71±0.18**	**0.74±0.15**	0.20±0.09	0.15±0.06	241.19±123.67	39.80±20.41
	≥ 0.4	**0.78±0.17**	**0.72±0.18**	**0.77±0.16**	0.22±0.09	0.17±0.07	175.75±100.25	29.00±16.54
	≥ 0.5	**0.79±0.17**	**0.75±0.19**	**0.80±0.16**	0.23±0.11	0.20±0.08	142.30±82.73	23.48±13.65
	≥ 0.6	**0.84±0.18**	**0.78±0.19**	**0.81±0.16**	0.26±0.11	0.23±0.09	130.10±71.75	21.47±11.84
	≥ 0.7	**0.86±0.15**	**0.80±0.17**	**0.84±0.15**	0.28±0.11	0.24±0.09	122.89±65.83	20.28±10.86
	≥ 0.8	**0.88±0.15**	**0.81±0.17**	**0.87±0.15**	0.34±0.13	0.26±0.09	116.80±61.90	19.27±10.21
Min-CQP-60	≥ 0.0	0.46±0.15	0.41±0.16	0.50±0.15	0.14±0.05	0.04±0.02	606.00±168.75	100.00±27.85
	≥ 0.05	**0.76±0.22**	**0.80±0.22**	**0.81±0.21**	0.25±0.11	0.29±0.11	158.57±98.42	26.17±16.24
	≥ 0.1	**0.85±0.23**	**0.79±0.24**	**0.87±0.22**	0.27±0.12	0.33±0.12	120.26±80.84	19.84±13.34
	≥ 0.15	**0.84±0.22**	**0.78±0.22**	**0.87±0.21**	0.33±0.13	0.36±0.11	104.35±65.04	17.22±10.73
	≥ 0.2	**0.86±0.19**	**0.82±0.21**	**0.88±0.18**	0.34±0.16	0.38±0.11	102.19±63.56	16.86±10.49

Third, we observe that just removing the irrecoverable PPG segments (i.e., data quality threshold of 0) already provides a fairly high correlation (≈ 0.7) [215] for the first order features in the lab environment. However, in the field, a data quality threshold of 0.3 is required to provide a similar level of correlation. Fourth, we note that the correlation profiles of the first-order features are all quite similar to each other. The correlation profiles for the HRV features are also similar to each other, but they are significantly lower than for the first-order features. This can be explained by the fact that first-order features have greater statistical robustness. Also, accurate computation of HRV features usually requires a high sampling rate. As compared to ECG that is sampled at 100 Hz in the lab and 64 Hz in the field, PPG is sampled only at 25 Hz due to battery limitations. Fifth, in the field environment, the correlation for HRV features

are 0.36 at best (for Mean-CQP-60), at which point, the yield is quite low. In the next section, we present a case study exploring the end-to-end application of the CQP-60 index to a more challenging problem: improving the robustness of stress inference from PPG data.

4.8 Deep Integration of the CQP Model to Improve the Robustness of PPG-Based Stress Inference

In this section, we present a case study leveraging the CQP model to improve the robustness of stress inference from PPG data. We again focus on assessing an accuracy versus yield trade-off using either ground truth stress labels (in the lab setting) or concordance with an ECG-based stress inference model (in the field setting). We consider two ways to leverage the CQP model. First, similar to past work, we consider thresholding the CQP-60 data quality index at different levels to expose a basic accuracy versus yield trade-off for stress prediction at the minute level. Second, we consider leveraging the CQP-5 data quality index to provide quality-weighted features in addition to quality thresholding using CQP-60. We begin by describing the stress inference process based on ECG data, where we follow the approach of [15]. We next describe the pre-processing, feature extraction and learning steps for the PPG-based stress inference. We conclude by describing the results of an experimental evaluation of the resulting models.

4.8.1 Stress Inference from ECG Data

For the ECG stress inference pipeline, we follow the steps from [15]. ECG data is first assessed for signal quality to identify acceptable segments of ECG. ECG conforms to a unique PQRST morphology and thus acceptable ECG segments are easily distinguishable from non signal components using simple thresholds on signal variance and range [216]. We detect R peaks via the widely-used Pan Tompkins algorithm [217]. We then use the criterion beat difference (CBD)-based method from [218] to filter out outliers and generate a clean RR interval time-series. Stress inference are computed for

every one-minute window of RR interval data [15, 219]. We consider a one minute window of RR interval timeseries usable if at least half of the window has acceptable ECG data and sufficient ECG RR intervals are present, following [15].

4.8.2 PPG Data Cleaning and Computation of Inter-beat Intervals

As described in Section 4.7, for PPG data we apply a bandpass filter to remove high-frequency noise, normalize the signal to remove rapid variability in signal amplitude, segment the signals into 5-second segments with 50% overlap, identify and remove irrecoverable segments, apply the CQP-5 model to obtain signal quality likelihood for all remaining segments, select the channel with the highest signal quality to represent the current segment, and compute the mean of these signal quality likelihood values in a minute to estimate the signal quality index (CQP-60) if at least 50% of the segments are present in the minute. For inter-beat interval computation for each minute, we compute the mean RR interval in each 5-second segment, using the spectral peak detection method from [82].

4.8.3 Quality-Integrated PPG Feature Computation

To train a machine learning model for inferring stress, we experiment with the 11 features used in [15], all of which are calculated from RR intervals over one-minute windows. The features include mean, median, 80^{th} percentile, 20^{th} percentile, variance, quartile deviation, low frequency energy (0.1–0.2Hz), medium frequency energy (0.2–0.3Hz), high frequency energy (0.3–0.4Hz) and low to high frequency energy ratio computed from RR intervals and heart-rate. Results from Section 4.7.8 show that estimation error of mean RR interval values is lower for higher values of the CQP-5 indicator. We use this finding to integrate quality weighting into the feature computations needed for stress inference.

Let r_i represent the mean RR interval and q_i be the signal quality likelihood of the i^{th} 5-second data segment in a minute of PPG data with with $n \leq 23$ such segments. q_i refers to the *CQP-5* value for i^{th} segment. We then compute the weighted mean RR

interval in a minute as $\bar{r}_w = \sum_{i=1}^{n} q_i r_i / \sum_{i=1}^{n} q_i$. Similarly, we compute the Weighted sample variance as $s_w^2 = \sum_{i=1}^{n} q_i (r_i - \bar{r}_w)^2 / \frac{(n-1)}{n} \sum_{i=1}^{n} q_i$. To compute percentile based features such as the median, interquartile range, 80^{th} percentile, 20^{th} percentile, we use the q_i values as frequency weights. To calculate weighted percentiles, we first normalize the data quality weights within each window according to $q_i' = q_i / \sum_{i=1}^{n} q_i$. Next, we sort the mean RR interval timeseries r_i. The weighted c^{th} percentile is given by the element r_k which satisfies $\sum_{i=1}^{k-1} q_i' \leq c/100$ and $\sum_{i=k+1}^{n} q_i' \leq c/100$. The heart rate per minute is computed from the inverse of weighted median RR interval (i.e., $c = 50$).

For calculation of frequency domain features, we use the Lomb-Scargle periodogram [220, 221]. Computation of spectral density using the Lomb-Scargle method requires temporal consistency to be preserved in the input time series. To introduce weights based on signal quality likelihood, we transform the mean RR interval timeseries values r_i using an exponentially weighted moving average. Using an exponential moving average with weighting, the RR interval value at the i^{th} time is computed as shown below. The periodogram-based features are calculated from the r_i^{ewma} timeseries using $\alpha = 0.7$.

$$r_i^{ewma} = \alpha[r_i q_i + (1-\alpha) r_{i-1} q_{i-1} + (1-\alpha)^2 r_{i-2} q_{i-2} ... + (1-\alpha)^{i-1} r_1 q_1] / \sum_{k=1}^{i} q_k. \quad (4.3)$$

4.8.4 PPG Feature Normalization to Account for Between-Person Variability

An important step in stress modelling is normalization of heart beat interval timeseries for minimizing the effect of between-person variability. Specifically, each person has a different baseline resting heart rate as well as different average heart rate variability features. Thus, to make the model more person independent, the authors in [219] proposed person specific normalization of RR interval timeseries. They transform the RR interval timeseries of each participant by converting them to z-score

values with the mean and standard deviation calculated from the baseline rest period of each participant. This procedure significantly improved the accuracy of stress model and later works on stress detection from ECG/PPG/Respiration have all incorporated this form of pre-processing [15, 103, 105]. We emphasize that z-score based standardization is a linear transformation and can only minimize the person-specific differences in the time-domain. Frequency domain HRV features calculated from RR interval z-scores will have the same value as with raw RR interval timeseries. To circumvent this, we propose person specific feature standardization instead. Let $\mathbf{F}$ be the feature matrix for stress model training calculated from a single person. $\mathbf{F}$ has a shape of $n \times 11$, with F_i^j indicating the value of j^{th} feature at i^{th} minute.

We normalize the columns of the feature matrix $\mathbf{F^j}$ as shown below, where $\hat{F}_{mean}$ is the mean feature value and $\hat{F}_{std}$ is the standard deviation calculated from the minutes belonging to baseline of the participant.

$$\mathbf{F^j_{normalized}} = (\mathbf{F^j} - \hat{F}_{mean})/\hat{F}_{std}, \tag{4.4}$$

To ensure we have sufficient data to compute the baseline, we use only those days from field data of a participant that has at least 60 minutes of acceptable data. As the lab study includes explicit baseline sessions before and after the stress sessions, we do not need any such criteria for the lab data.

4.8.5 PPG Stress Model Training

Using the normalized features from the lab data, we train a support vector machine with radial basis function kernel to optimize the leave one subject out cross validation score. For unambiguous labels of stress and no stress, we use the lab stress protocol sessions as shown in Figure 4.2a. As introduced in [219], all data when the participant is undergoing a stress task is labeled as stress and the rest and recovery sessions as no-stress. As these stress tasks are validated by psychologists to produce a

physiological stress response, they provide validated ground truth labels for stress. The choice of support vector machine for modeling stress stems from its ability to learn accurate models with a limited number of features while avoiding over-fitting using regularization. Using Platt's scaling [222], the model outputs stress/not-stress class probabilities for each minute of data. We choose the F1-score to be our performance metric as the classes are unbalanced. All hyper-parameters are optimized using grid search to maximize classification performance.

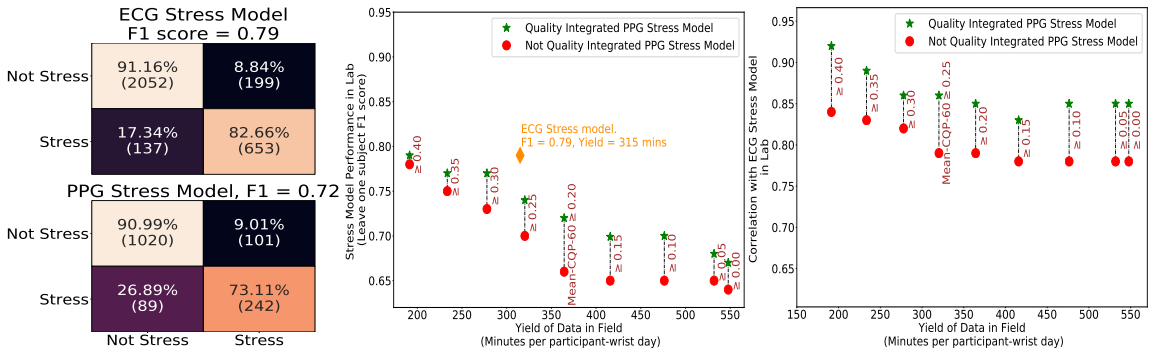
4.8.6 Experiments on Accuracy vs. Yield of Stress Inference in the Lab

We begin by reporting the classification performance of the PPG and ECG stress models on labeled data from the lab similar to existing works [15, 219]. We find that the ECG-based model achieves a test F1 score of 0.79. For the PPG-based model, we select a quality threshold of Mean-CQP-60 ≥ 0.2. At this quality level, the PPG-based stress inference model achieves a test F1 of 0.72 using data from the left wrist only, 0.69 using data from the right wrist only, and 0.7 when pooling data from both wrists. This shows that stress inference computed from the left wrist is slightly more accurate than stress inference computed from the right wrist. Figure 4(a) shows the confusion matrix for the ECG and PPG models where the PPG results are for the left wrist only with Mean-CQP-60 ≥ 0.2 as the quality threshold (as noted above). As we can see, the difference in F1 score between the ECG and PPG models is largely due to an increased number of false negatives in the PPG-based model (27% versus 17%).

Table 4.6: Yield breakdown on field data using Mean-CQP-60 ≥ 0.2

	ECG Based Model	PPG Based Model		
		Left Wrist Only	Right Wrist Only	Using Either Wrist
No. of Participant-Days in Field with Acceptable Data	738	894	886	978
Amount of Acceptable Data (mins/day in field)	375.49±199.91	387.97±149.11	380.28±148.07	384.16±148.62
Data Usable for Stress Assessment (mins/day in field)	315.06±197.16	369.15±148.29	359.36±147.89	364.28±148.13
Total Data for Stress Assessment (mins) in field	232,560 (=3,876 hours)	330,020 (=5,500 hours)	318,392 (=5,306 hours)	356,265 (=5,937 hours)

Importantly, we note that the accuracy of PPG-based stress inference in the lab is better than that reported in prior work, even at the data quality threshold of 0.2. For



(a) (b) (c)

Fig. 4.5: Yield-Accuracy Trade-off of Quality Aware vs. not aware stress modelling. Figure(a) shows the Confusion Matrices for ECG stress model and PPG Based Stress Model from left wrist. Figure(b) shows the Leave one subject out cross validation scores of PPG stress model developed on data from left wrist in the lab as a function of increasing minimum thresholds on Mean-CQP-60. Figure(c) shows data yield vs correlation with ECG stress model of both quality integrated and not integrated PPG based stress model in lab. Green represents quality integrated model whereas red represents quality not integrated model. X-axis in both (b) & (c) shows the mean number of minutes available in field for corresponding thresholds on Mean-CQP-60

stress inference from wrist-worn PPG sensors (supplemented with electrodermal and inertial sensors) using a lab stress protocol consisting of mental arithmetic task of increasing difficulty, [105] reported an F1 score of 0.67 for their best classifier for one minute data segments. For comparison, using only 11 features on PPG data, without using eletrodermal or inertial sensors, we obtain an F1 score of 0.72, which as we show in Figure 4.5b can be improved by further by raising the data quality threshold.

Table 4.6 provides a breakdown of the yield for the ECG-based model and the PPG-based model using Mean-CQP-60 ≥ 0.2 when applied to field data. As described in [15], data that is affected by torso motion is not usable for stress inference. Hence, all data that is collected during significant torso motion as detected by the accelerometer in the chest-worn sensor is excluded from stress inference. As we can see, the left wrist also provides a somewhat higher data yield than the right wrist, while both provide significantly higher yield than ECG based on the quality threshold used. The improvement in yield with the left wrist may be because the right wrist is dominant for

the vast majority of participants and hence is more likely to be affected by motion and other artifacts. Although we are only able to assess left vs. right from our labeling of the wrist sensors, we hypothesize that the non-dominant wrist can provide even better accuracy and yield.

Next, we consider the performance of the PPG stress model using both quality thresholding and quality integrated feature computations. As we can see in Figure 4(b), the quality integrated computation out-performs the non-quality integrated computation for all quality thresholds. In this plot, the horizontal axis shows the yield of each quality threshold when applied to field data, while the vertical axis shows the corresponding test F1 score. These results can be viewed either as providing significantly higher classification performance at the same yield, or as providing significantly higher yield at the same level of classification performance. Lastly, Figure 4(c) shows the correlation between the quality integrated and non-quality integrated PPG-based stress inference model and the ECG-based model at different CQP-60 quality thresholds. As we can see, the higher quality thresholds result in increased correlation between the models. In addition, the quality-integrated model uniformly out-performs the non-quality integrated model at all quality thresholds, providing a substantially improved correlation-yield trade-off.

4.8.7 Experiments on Accuracy vs. Yield of Stress Inference in the Field

In the field setting, we do not have labels for stress, thus our predictive analysis focuses on comparing the ECG-based model to the PPG-based models in terms of Pearson correlation. As we saw in the previous section, the ECG-based model has strong performance in the lab in terms of F1 score, thus convergent validity of ECG and PPG-based models is a reasonable evaluation procedure.

We present the performance of the PPG stress model using both quality thresholding and quality integrated feature computations in the field in Table 6. As we can see, the correlation between the ECG and PPG-based models again increases as the

Table 4.7: Correlation Between ECG and PPG-based Stress Models and Yield of total data in Field with and without Quality-Integration

	Pearson Correlation with ECG Stress Model in Field		Yield of Usable Data in Field	
Minimum Threshold on Mean-CQP-60	Quality Integrated PPG Stress Model	Not Quality Integrated PPG Stress Model	Total Data(hours)	Percentage Relative to ECG (3876 Hours)
$\geq$0.0	0.48±0.21	0.42±0.23	8421.38	217.27%
$\geq$0.05	0.49±0.22	0.44±0.23	8158.45	210.49%
$\geq$0.1	0.53±0.23	0.49±0.24	7293.17	188.16%
$\geq$0.15	0.60±0.22	0.51±0.24	6324.88	163.18%
$\geq$0.2	0.63±0.20	0.55±0.24	5500.35	141.91%
$\geq$0.25	0.69±0.18	0.62±0.24	4710.27	121.52%
$\geq$0.30	0.70±0.18	0.63±0.23	3899.48	100.61%
$\geq$0.35	0.74±0.18	0.64±0.24	3122.03	80.55%
$\geq$0.4	0.76±0.18	0.64±0.24	2408.28	62.13%
$\geq$0.45	0.76±0.19	0.65±0.27	1766.20	45.57%
$\geq$0.5	0.77±0.23	0.68±0.28	1284.22	33.13%
$\geq$0.55	0.77±0.20	0.68±0.24	992.75	25.61%
$\geq$0.60	0.78±0.23	0.70±0.25	803.53	20.73%

CQP-60 quality threshold is increased. In addition, the use of quality-integrated feature computations results in a substantial improvement in correlation at the same level of yield compared to the non-quality integrated model. We note that when operating the PPG-based model with a yield that matches that of the ECG-based model, we obtain a correlation of 0.7 using the quality-integrated stress model. On the other hand, the not-quality integrated stress model achieves this level of correlation at the quality threshold of 0.6, for which the data yield is only 21% of that obtained from ECG. Lastly, we note that the correlation we obtain for stress inference after integrating quality is comparable to what we obtain for first-order features such as heart rate, 80^{th} percentile, and 20^{th} percentile (see Table 4.4). For example, for the 0.2 quality threshold, we get a correlation of 0.63, which is comparable to the correlation of 0.61-0.62 for the first order features and significantly better than 0.16-0.18 we obtain for the second order features. This means that our approach is able to effectively overcome some of the apparent weakness in the base features.

4.9 Limitations and Future Works

This work has several limitations that open up the opportunity for future research. First, we used 5 second segments as a unit for data quality assessment and low-level feature computations. Future work can investigate different choices for this window to see how they impact the accuracy of features and inferences derived.

Second, this work used a single device set at a sampling frequency of 25 Hz for PPG data collection in both the lab and field settings. Future work can investigate the impact of sampling frequency, types of PPG sensors, and different channels of PPG data on the quality of data and consequently on the accuracy of features and inferences.

Third, this work presented a method to estimate and represent the quality of PPG data and showed how to incorporate quality into machine learning models to improve the robustness of inferences. However, there are several sources of uncertainty in feature computations. Future work can investigate comprehensive approaches to estimate uncertainty, represent it succinctly, and propagate it in the processing pipeline.

Fourth, for the purpose of assessing activity confounds in the context of stress inference, we leveraged the accelerometer in a chest-worn sensor. Future work can develop methods to make an assessment of physical activity confounds from wrist-worn sensors which capture the movement of wrists, instead of or in addition to capturing torso motion.

Fifth, this work assessed the concordance of stress inference in the field estimated using PPG-based models with that output by an ECG-based model using paired ECG and PPG data. Future work can assess the concordance of quality-informed stress inference with self-report collected in field.

Finally, in addition to stress, several other physiological states can potentially be inferred from PPG sensors such as pain, craving, and drug use. Future work can investigate the integration of data quality into models for these domains and evaluate its ability to improve accuracy-yield trade-offs.

4.10 Chapter Summary

This work showed that assessment of wrist motion may not be a sufficient indicator of PPG sensor data quality. This should be expected as PPG data quality may be affected by several other factors including loose attachment and contamination from ambient lighting. We have instead proposed an approach to estimating PPG data quality using supervised learning and shown how the resulting continuous data quality indicator, CQP, can be more deeply integrated into subsequent inferences to significantly improve accuracy-yield trade-offs both for the computation of individual features and for complex high-level inferences (e.g., stress).

As new research seeks to infer stress, pain, craving, drug use, and other physiological states and events from PPG sensors that are now integral components in smartwatches and activity trackers, this work provides a new approach to improve the yield and accuracy of these computations in real-world settings. This improved accuracy and coverage may provide higher quality inputs to down-stream processes for a variety of applications in wellness, self-care, and precision health care. For example, our approach could be leveraged in the context of just-in-time adaptive interventions to increase the number of time points and the contexts where required inferences can be accurately computed to support intervention decisions, helping to lead to overall improved therapeutic efficacy.

Chapter 5

rSmoke: Orientation-Invariant Detection of Smoking Events from Wrist-worn Inertial Sensors

5.1 Introduction

In Chapter 3, we describe the machine learning-based deep neural network models to continuously output the imminent risk of smoking lapse using mobile sensor data collected in the natural environment. We first employed state-of-the-art mHealth prediction models to passively detect varying risk factors dynamically. These factors represent the physiological (e.g., stress), behavioral (e.g., activity), and environmental (e.g., proximity to smoking spot) contexts of participants. Using the continuous estimate of these risk factors as input, we trained LSTM-based deep models called *mRisk* (See Chapter 3) to predict the risk of smoking lapse. Our primary goal is to facilitate the effective design and delivery of just-in-time smoking interventions based on the risk score produced by our model. Hence, we evaluated the *mRisk* models by simulating their ability to inform the delivery of just-in-time interventions. However, the effective deployment of a continuous smoking lapse risk estimation model to deliver just-in-time adaptive interventions requires addressing multiple outstanding challenges of the *mRisk* modeling pipeline.

In *mRisk* (see Chapter 3), we employed chest-based ECG, Respiration, and inertial motion sensors for passive sensing of human health and wellness states. Owing to the inconvenience of wearing chest-wrapped devices daily, the practical utility of using such sensors outside of academic use remains limited. Researchers in [8] note that the "complexity of equipment" reduces the quality of collected data in the smoking cessation field study employing chest-worn *AutoSense* sensor suite [72], thus limiting the feasibility of intervention design based on wearable sensing alone. On the other hand, wrist-worn wearables or sensor-fitted smartwatches have seen growing adoption in research and commercial use due to their convenience and the ability for continuous

monitoring. Adapting our models and methodologies to work with wrist-worn wearable sensors alone will significantly boost the practical utility of deployment in the wild.

The first step in achieving this goal is to enable continuous inference of dynamic risk factors from wrist-worn sensor data. We took the first step in this direction by developing *CQP* (see Chapter 4). In *CQP*, we proposed methods to enable continuous stress inference from wrist-worn sensor data. The next significant challenge in adopting a wrist-only smoking lapse risk estimation model is our ability to employ a smoking detection model using wrist-worn sensors alone. For *mRisk*, we used *puffMarker* [10] to detect smoking occurrences passively.

puffMarker uses wrist-worn inertial motion sensors and chest-worn respiration sensors to detect smoking puffs. A collection of these detected puffs in close temporal vicinity with each other together makes up a smoking episode. The smoking detection model is essential for developing and deploying risk estimation models. First, we need the precise timing of smoking lapses for training the *mRisk* models. Smoking self-reports corroborate these smoking lapse detections through Ecological Momentary Assessments (EMAs). One of the important limitations of the *mRisk* models is the lack of enough positive (high-risk) instances. We had 84 confirmed lapses from 56 participants. Thus, the number of positive cases in which the participants were confirmed to be at high risk of smoking lapse was limited in scope. Second, we need the timing of smoking instances in specific locations to extract the personal smoking spots. We computed the smoking spots using location data and confirmed smoking instances from the pre-quit period. These spots are used for computing features related to location proximity to smoking spots and visitations. Therefore, we need a smoking event detection model from wrist-worn inertial sensors alone to enable real-time risk estimation using wrist-worn sensors.

In this chapter, we develop an orientation-invariant smoking detection model using wrist-worn accelerometers and gyroscope sensors. We utilize the video-coded

labeled data from participants' natural environments to train and test our modeling strategy. This is the first time real-life field data has been used to develop smoking detection models from wrist worn sensor. We identify and address the critical challenges specific to orientation of wrist-worn sensors in the natural environment. These challenges include variability in sensor configurations, variability in axes orientation due to sensor placement along with lack of sufficient field collected training data in deploying a smoking event detection model in the field. We developed novel strategies to construct smoking events from the noisy detected puffs and distinguish these events from the non smoking ones. We devised a novel strategy to incorporate sub episodes within candidate smoking events for smoking event prediction to overcome the paucity of few candidate events. We evaluate our developed methodology with field data from 200+ participants from two different studies with different sensor configurations.

Our results show, the proposed sensor orientation invariant *rSmoke* model obtains a precision of at least 0.65% in two test studies with original and switched sensor mounts. *rSmoke* model also outperforms the *puffMarker* model in the percentage of the EMA reported blocks with smoking correspondence. *rSmoke* provide increased recall of smoking self-reports, hence providing us with improved data coverage for training smoking risk estimation model.

5.2 Robustness Challenges to Smoking Detection using Wearable Wrist-worn Sensors in the Field

We focus on developing a smoking detection model using wrist-worn inertial sensors. Second, we train and test a real-time smoking risk prediction model using the wrist-based smoking detection model from data collected using wrist-worn sensors alone. Both these steps present several technical challenges. We outline these challenges below in distinct categories.

(a) Wrist coordinate system showing three general orthogonal direction axes.

(b) Sensor Mounting on Different Devices.

(c) Lateral axis aligned to gravity.

(d) Lateral axis opposite to gravity.

Fig. 5.1: Figures showing sensor mounting and axes orientation variability.(Figure (b) taken from [1])

Variability in Sensor Configurations

Figure 5.1a shows the general wrist coordinate system with three mutually orthogonal axes as defined by researchers in [1]. The lateral axis (l) is aligned with the arm, the perpendicular axis (p) is aligned with the thumb, and the vertical axis (v) is the gravity axis when the palm is parallel to the earth's surface [1]. Although we define these three as the general directions, accelerometer sensors mounted on devices provide data using the Cartesian coordinates - (x, y, z). Usually, the z-axis corresponds to the vertical axis. However, the x and y axes mounting can vary in different sensors. Figure 5.1b shows two different 3-axis accelerometer sensors with two configurations - $(l, p, v) = (x, y, z)$ and $(l, p, v) = (y, x, z)$. The same sensor can change its wrist configuration owing to software or firmware updates. Hence given inertial sensor data from the field as Cartesian coordinates, we have to identify the sensor configurations in the form of the general wrist coordinate system.

Variability in Axes Orientation Owing to Sensor Placement

Figures 5.1c and 5.1d show two different orientations of the inertial sensor on the wrist owing to participants' slightly different placements on hand. In both these placements, the configuration of the sensors is $(l, p, v) = (x, y, z)$. However, the lateral

axis ($l = x$) points away from the arm in Figure 5.1c and the opposite in Figure 5.1d. We also observe the same phenomena for the perpendicular axis ($p = y$). This phenomenon is ubiquitous since individuals wear wristwatches in their unique way. Placements can also vary between days and periods for a single individual. Hence, for smoking detection using wrist-worn inertial sensors, we must identify the directions of inertial movement along each axis as measured by the sensors.

Lack of Sufficient Training Data from Natural Environment

Developing a working smoking detection model to deploy in the field requires a comprehensive dataset containing the influence of sensors, individuals, context, ground-truth labeling, and other factors. The dataset must be large enough and incorporate enough diversity to account for the many differences that can arise in the field environment. Distinguishing smoking from similar behaviors involving hand-to-mouth gestures, such as eating or drinking, requires precise labeling of the smoking puff events. This constrains the individual participants to wear video collection devices that the annotators can use to label the different events. Given these limitations, we need to maximize the efficiency of our modeling scheme to extract maximum value from the limited amount of information available to learn from.

5.3 Prior Works on Smoking Detection and Our Contributions

Works on smoking puff detection using wrist-worn inertial motion sensors mainly focus on detecting the hand-to-mouth gestures of smoking puffs. Inertial motion sensor units have been widely used for assessing daily life activities. Previously published studies typically involve collecting lab data from participants with one or more sensors placed into a reference position [57, 58, 59, 60, 61, 55, 56]. Developing smoking detection model with data collected from lab supervision limits the utility of the developed model in the field. Several studies have collected data from the natural field environment [58, 10, 62]. However, they also assume a reference position for the smart-watch-based inertial sensors and do not address the variability owing to sensor

mounting, placement, and other factors. Researchers have used other sensors to estimate the smoking puff action accurately. Respiration sensors (RIP) have been used alone [55, 56] or in combination with inertial accelerometry [10] to identify smoking puffs. RIP and RF hand-to-mouth proximity sensors [115] has been used to detect and characterize cigarette smoke inhalations. Respiration sensors are chest-worn and increase the burden on participants to wear a chest-belt device in their daily lives. In [61], authors mounted inertial sensing units inside a smart lighter to better distinguish the smoking puffs. In [58], researchers advocated using 9-axis IMU units containing quaternions to more accurately estimate the trajectory of hand motion. [116] deployed a chest-worn thermal-sensing wearable system that captures spatial, temporal, and thermal information around cigarettes and the wearer to passively detect smoking events throughout the day. Works involving novel sensing schemes to detect smoking behaviors are ongoing, and the unique challenges facing them are independent of the ones affecting wrist-worn sensors. The ever-growing adoption of inexpensive wrist-based wearables in smartwatches and related edge devices attests to the large impact of a smoking detection model developed from wrist sensors alone. In contrast, we aim to build a smoking detection model using data from 6-axis Inertial motion units (3-axis accelerometer and 3-axis gyroscope) in the natural environment. These ubiquitous sensors come integrated within commodity smartwatches. An accurate, inexpensive, and working smoking detection model using these sensors will be a significant leap in the research on detecting smoking detection. Our work builds upon *puffMarker*, which used 6-D IMU and Respiration sensors to detect smoking puffs and identify smoking episodes using data collected from the field environment with human supervision. However, our work is unique in many ways.

First, we do not use respiration sensors. We want to ensure maximum practical utility for our developed model by only using wrist-worn inertial sensors for smoking detection in the natural field environment.

Second, the authors collected the training data from smokers with a human observer present. The observer marked the start and end of the smoking puffs and coordinated the sensor placement and other factors. We build our training dataset using data collected from participants living their daily lives in the natural environment. Participants wore a video collection device that recorded their daily life activities. Later, human annotators annotated the smoking puffs from the videos collected. The presence of an accompanying human observer significantly impacts the data collection environment and limits the degrees of freedom and diversity of the collected data.

Third, *puffMarker* assumes a fixed direction of the inertial motion axes - Y axis opposite to gravity. The authors do not address the scenarios of axis identification in the case of different types of sensor mountings (Figure 5.1b). In contrast, we propose a basic rule-based approach to identify which axes are aligned with the lateral and perpendicular directions, thus making our methodology robust to any change of sensor types and configuration changes. Also, since *puffMarker* assumes a reference set of directions for all three axes in both hands, it is not robust to the different types of sensor placements resulting in different orientations. In [10], authors detect a positive rise of the accelerometer y-axis time series to designate candidate puffs. However, in case of a change in orientation resulting from different sensor placement, the positive rise pattern will change to a fall in the negative direction resulting in a failure case. Thus, we do not employ individual axis time series to identify the candidate puffs. We utilize the gyroscope magnitude time series to identify candidate puffs. The magnitude time series is invariant to the direction of the individual inertial motion axes. Next, we propose rule-based ways to identify the direction of the lateral axis of the accelerometer and align it to gravity before feature computation and model building.

Finally, *puffMarker* focuses on smoking puff detection using a Support vector machine (SVM) based machine learning model. Our results show that puff detection using field-collected data is not sufficiently accurate in the field environment. Hence, we

extend the output of our trained puff detection model to construct smoking episodes using another machine learning-based model. The advantage of using a probabilistic smoking event detection classifier is the ability to represent the whole smoking episode using top-level features representing information about multiple puffs at a time. Also, in our use case of smoking lapse detection, we corroborate the smoking events given by the model using self-reports. The confirmation through EMA-provided self-reports reduces the chances of false positives. Our proposed probabilistic smoking event classifier allows us to select the appropriate operating threshold for capturing a significant portion of smoking lapse behaviors within the study.

5.4 Dataset Description

We develop the *rSmoke* model using data collected from participants in the natural environment as part of a smoking cessation research study. We evaluate our methodology using data from two other test studies. In both these studies, participants completed regular and random Ecological Momentary Assessments (EMAs), where they self-reported their recent smoking events. We apply *rSmoke* and report the accuracy of detecting these self-reported smoking events. The Institutional Review Board (IRB) approved all the studies. We now describe them in the necessary details.

5.4.1 Training Data for Smoking Detection

We have training data from 10 daily smokers in the field. The participants were part of a smoking cessation research study and wore the chest-worn AutoSense [72] and wristbands on both wrists. We develop *rSmoke* using the 6-axis accelerometer and gyroscope sensors fitted within the wristbands. The sampling frequency is 16.33 Hz for both the accelerometer and gyroscope data. Participants also wore a video collection device (goPro). We annotate the videos to designate ground truth labels of smoking puff gestures. Participants provided smoking data in their pre-quit period, where they wore the sensor suites daily. We have data from 15 user days with 360 labeled smoking

puffs from 41 smoking events. Out of the labeled smoking puffs, 263 are smoking with their right hand, 97 with their left hand.

5.4.2 Smoking Data from Field With Original Sensor Mount

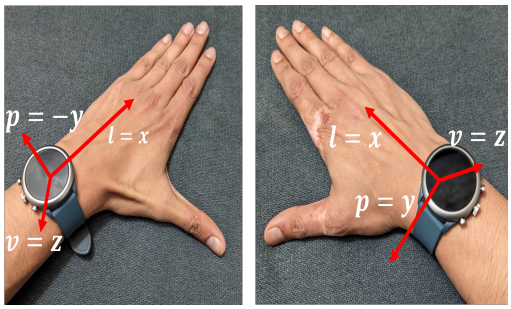
This is the same study described in Chapter 3. Participants wore a chest-worn AutoSense sensor suite fitted with ECG, Respiration, and Accelerometry sensors. Participants also wore wrist-based smartwatches on both wrists. The sensor suites worn in this study are identical to the training study with the wristband collecting 6 axis accelerometer and gyroscope data at a sampling frequency of 16.33 Hz. We have data from 92 participants in the field wearing both the chest and wrist sensor suite. Alongside the sensor data, the participants also completed $3,719$ EMAs. We use the smoking self reports in EMAs to evaluate our smoking detection model.

5.4.3 Smoking Data from Field With Switched Sensor Mount

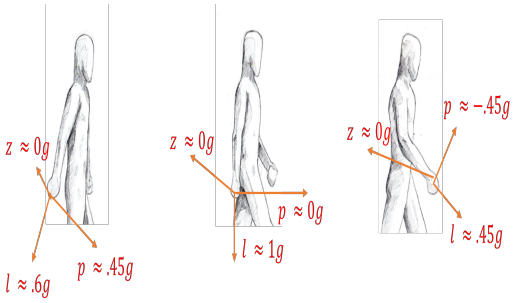
We use the same study from Chapter 4. Similar to the previous test study, this study also involves daily smokers in a smoking cessation research. Participants wore the chest-worn AutoSense sensor suite for chest-sensing and smartwatches on both wrists fitted with a PPG sensor and 6-axis accelerometer and gyroscope sensors. The sampling frequency of wrist-worn PPG, Accelerometer and Gyroscope sensors was 25 Hz. We have chest and wrist-sensor data from 97 participants with 1884 EMAs completed. The inertial motion sensor used in this study is different from the one used in the training or first testing study. The IMU sensor configuration is switched in this study. We propose methods to identify and align the sensor configurations in both the testing studies.

5.5 Inertial Sensor Mount Identification and Axis Alignment

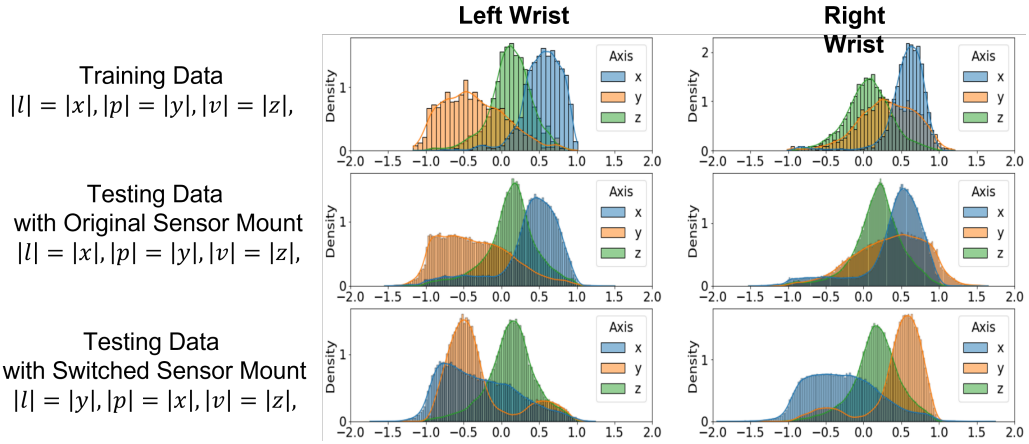
Wrist-worn inertial motion sensors can be worn in multiple ways on wrists. Also, different sensors can have different configurations of the inertial axes of a movement, contributing to a high diversity and differences in data distribution. We must account for these differences in our methodology for detecting smoking events using wrist-worn motion sensors. Typically studies involving wrist-based IMU sensors assume a fixed



(a) Axis Orientation of 3-axis accelerometer in both wrists.



(b) Variability in lateral and perpendicular Axes during walking.



(c) Distribution of accelerometer axes in both wrists during walking.

Fig. 5.2: Figures showing (a) opposing orientation of the perpendicular axis between left and right wrist, (b) three different walking moments showing values of lateral and perpendicular axis in the right wrist (human figurine copied from [2]), (c) Distribution of the accelerometer axes values during walking in our studies

configuration of the 3 orthogonal axes of movement. They ensure this compliance in their data collection process by constraining their data collection environment to a lab or controlled setting.

Existing works on orientation invariant processing of inertial wrist sensor data involve supervised classification [1, 121], incorporating additional sensing medium [119, 122], and transformation of tri-axial sensor values to an orientation-invariant representation [123, 124, 125, 126, 127, 128, 129, 130, 117, 131]. When transformed into a different system of coordinates, the individual sensor axes are each changed to a different sequence with loss of the initial meaning and information. In

proposing an orientation-invariant method of smoking detection, we aim to preserve and utilize the already known patterns present in the original sensor signals. Our orientation in-variance approach tackles two distinct problems. We propose methods to identify the configurations of a given wrist-worn sensor using the distribution of the accelerometer sensor signals during moments of walking. More formally, given a 3-axis accelerometer sensor with x,y, and z axis, we must determine their orientation in the general wrist coordinate system of lateral, perpendicular, and vertical (l, p, v) axes shown in Figure 5.1a. Next, we propose to align the sensor axis corresponding to the gravity line in real time. Our proposed methods address the orientation-related challenges that affect smoking detection from wrist-worn inertial sensors in the natural field environment.

In this section, we describe our proposed methodology to achieve these goals. We employ the convolutional neural network-based activity detection models proposed in Chapter 4 (see Section 4.2). Recall that we detect five different physical activity labels in 20-second windows using the magnitude of accelerometer data - Walking, Stairs, Stationery, Exercise, and Sports. We apply our activity detection model to the data collected and extract the times of various physical activities. Next, we concentrate on the moments when an individual walks while wearing the wrist-worn accelerometer sensors. The key idea is to utilize the distribution accelerometer signals during walking to devise a rule-based algorithm for identifying the three axes of inertial movement and the direction of the lateral axis.

5.5.1 Distribution of Vertical Axis during Walking

Figure 5.2b shows the direction of the vertical axis during walking at three different points within a walking step. With our hands pointed downwards during walking, the direction of the vertical axis remains parallel to the earth's surface. Thus, the vertical axis remains orthogonal to gravity's direction; most values will be close to zero. In all six (three studies, two wrists = six total) of the plots, the distribution of the z-axis shows this pattern. There is a peak centered on zero with an almost equal spread

in both negative and positive directions. The z-axis distribution plots are consistent with the expected distribution of the vertical axis. The general pattern of vertical axis distribution centering on zero allows us to determine the exact individual axis to be considered the vertical axis. Therefore, Figure 5.2c shows |v| = |z| for all three studies.

5.5.2 Key Idea: Distinguishing between Lateral and Perpendicular Axes Distribution during Walking

With the vertical axis identified, our next goal is understanding the perpendicular and lateral axes. Figure 5.2b shows three different points in a walking step while wearing a wrist-IMU sensor on the right wrist. Our drawing adopts the (l, p, v) coordinate system with the lateral axis aligned to the direction of the fingers and the perpendicular axis in the same direction as the thumb. We first explain the expected distribution of lateral and perpendicular axes at moments of walking. Understanding their individual and pairwise interaction will allow us to construct rules for accurately identifying the lateral and perpendicular axes.

Distribution of Lateral Axis

Concerning the lateral axis, we find two situations that may arise. First, an individual axis of the accelerometer (assume k) is aligned to the lateral axis ($l = k$, both pointing towards gravity). Figure 5.2b illustrates this scenario with the lateral axis pointing towards gravity. If we plot the distribution of the lateral axis value, we will see that most values are close to $1g$ in the positive direction. Second, in the case where the individual accelerometer axis is opposite to the lateral axis ($l = -k$). The significant distribution mass will be negative (close to $-1g$) for the later scenario. In both these scenarios, we expect the distribution to display a sharp peak with significantly less probability mass around the origin. Similarly, the location of this peak can also inform us as to the orientation of the inertial movement axis in the sensor to gravity.

Distribution of Perpendicular Axis

The distribution of the perpendicular axis during walking shows significant variability. Since the 'perpendicular' axis is aligned to the thumb, its general direction aligns with the swing direction of our hands during walking. This is true for both hands owing to the opposable thumbs. In Figure 5.2b, we see three different states of the perpendicular axis in the right hand during walking. The perpendicular axis shows a broader range from negative to positive g. Thus, we expect the distribution of the perpendicular axis to have more spread along the entire range of possible accelerometer values. However, unlike the vertical axis, the distribution of the perpendicular axis values will not be centered around zero. The reasoning behind this phenomenon involves two intertwined arguments. First, the longitudinal position of our hand relative to our body will determine the prominence of positive vs. negative values in the distribution of the perpendicular axis. If the wrist is behind the body line perpendicular axis will be positive. And if the wrist is in front of the body line, the perpendicular axis will be negative. However, the hand's position relative to our body is seldom symmetrical in a walking episode. Individuals can walk with their wrists in their pockets or hold a cup of beverage. Multiple such scenarios can disturb the symmetry of the perpendicular axis. Most of these scenarios involve the wrist in front of the body line. This indicates that the distribution will be shifted to one side. However, we can not definitively say which direction the shift will be. This is due to the second element of our argument. Considering Figure 5.2a, for the exact position of the sensor-fitted watch on our wrists, the perpendicular axis aligns with the accelerometer axis for the right hand. In contrast, it is precisely the opposite for the left hand. Even for a single wrist, wearing the watch in different positions will affect the direction of the accelerometer axis aligned to the perpendicular one. And finally, for a single wrist (assume right), the sensor can be configured so that even when one of the accelerometer axes is precisely aligned to the lateral axis pointing towards gravity, another accelerometer axis is not aligned to the

thumb direction. Thus, we can not be precise about the exact direction of the perpendicular axis. We can only say that the distribution will be spread along the whole range of values with either a positive or a negative skew present.

With the expected lateral and perpendicular axis patterns explained, we can now describe our approach toward identifying the sensor configuration and lateral axis alignment.

5.5.3 Identification of 3-Axis Inertial Sensor Configuration

Extending our arguments to Figure 5.2c, we can see that in both training and field study 1, the accelerometer x-axis shows a sharp peak in the positive direction for both left and right wrists with little to no probability mass around zero. Based on this, we can claim that the direction of the x-axis of the accelerometer in the training study and field study 1 is along the same straight line drawn by the lateral axis. We further strengthen our argument by explaining the distributions of the y-axis in both studies. As described beforehand, the distribution of the y-axis in both studies displays a broader range with a significant skew on either the positive or negative side. Considering these phenomena together, we can definitively say that for training study and field study 1, the x-axis corresponds to the direction of the lateral axis, and the y-axis corresponds to the direction spanned out by the perpendicular axis.

We employ a different inertial sensor for field study 2 compared to field study 1 and the training study. In field study 2, the accelerometer y-axis depicts sharp peaks in the distribution, and the x-axis seems more spread out along the entire horizontal range of values. This is opposite to the x and y axes distribution in the first two studies. For field study 2, we observe that the y-axis corresponds to the same pattern expected from the lateral axis of the wrist coordinate system. The y-axis points primarily toward gravity on the right wrist, and the y-axis points opposite gravity in most cases on the left wrist. This results in a sharp peak-like distribution for the y-axis for both left and right wrists. Interestingly, we see both negative and positive peaks in the distribution. The scant

opposite scenarios of the right-wrist y-axis pointing opposite to gravity and the left-wrist y-axis towards gravity also result in small peaks in the negative and positive directions, respectively. All the evidence points towards switching the x and y axes' directions in field study 2 compared to both studies. Our proposed methodology of investigating the distributions of the individual accelerometer axes allows us to find out this switching scenario. Also, this switching can happen during the study due to a change in sensor type or sensor configuration. It affects the robustness of any developed model on top of the inertial wrist-sensor data.

5.5.4 Investigating the possibility of Exact Alignment of Individual Accelerometer Axes

So far, we have proposed methods for the general identification of the 3-axis accelerometer sensor relative to the wrist-coordinate system. Our methodology for identifying the perpendicular and vertical axes does not provide the exact alignment direction. Based on the distributions of the accelerometer values during walking, we can only detect the accelerometer axis corresponding to the straight line spanned by the perpendicular or vertical axis. Both these axes have significant variability in wearing and assuming a fixed set of directions in the sensor (similar to *puffMarker* [10]) introduces noise in the modeling methodology. Thus, in our proposed smoking detection methods, we do not assume a fixed direction of the two accelerometer axes corresponding to the perpendicular and vertical axes. This limitation introduces significant noise in the pitch and yaw angle estimation. Therefore, we refrain from using pitch and yaw values for modeling. However, based on the distributions during walking, we can accurately identify the exact alignment of the accelerometer signal corresponding to the lateral axis. Recall the two scenarios of lateral axis distribution from Section 5.5.2. If the accelerometer axis' direction points towards gravity (similar to the lateral axis), the peak of the distribution during walking will be on the positive side (close to $1g$). And if it points away from gravity, the peak will be close to $-1g$. Using this information, we can

dynamically align the accelerometer axis toward gravity. We consider the distribution of each individual on a day-by-day basis, and based on the location of the peak in the distribution, we can precisely align the axis with the lateral one. Aligning the identified axis towards gravity allows consistent computation of the exact rotation around the lateral axis. Therefore, using our axis identification and alignment methodology, we improve the robustness of our models by first identifying the limits of the wrist-worn inertial sensors and second by focusing on the variables or features we can robustly compute after employing the proposed processing steps.

5.6 *rSmoke*: Smoking Episode Detection

This Section presents an overview of the *rSmoke* model. We extend the existing literature on smoking detection by adopting the iterative steps of first detecting individual smoking hand-to-mouth puff gestures using classical models and then using the inferred puffs to construct a smoking episode detection model. We only use the 3-axis accelerometer and 3-axis gyroscope time series in a single wrist as input signals. We first explain the data curation steps involved for wrist-worn inertial sensor data. Next, we present a detailed overview of the smoking puff detection model. Finally, we describe the *rSmoke* smoking episode construction and modeling strategies.

5.6.1 Data Preprocessing

The first step in processing the inertial sensor data is identifying the moments when the sensor was not worn on the wrists. We utilize the deviation in accelerometer signals to detect if the sensor was kept on a static surface compared to the wrist. When accelerometer sensors are kept stationary on a fixed surface, such as on a table or within a box, the signals have little to no deviation present. Worn on wrists, even when stationary, blood flow beneath our skin, and inhalation and exhalation of air create minuscule movements in the sensor signals. There have been works on heart rate and respiration detection by leveraging these minute movement patterns. For our use case, we employ a simple threshold-based approach to discard all the segments that do not

have this minimum deviation. We segment the accelerometer data into 5-second segments and compute the standard deviation of the magnitude. If the standard deviation is above a certain threshold, we keep the segment; otherwise, we discard it. This simple methodology is generalizable across different sensors and, without complex data processing requirements, allows us to discard segments with no information.

The next step in processing the inertial data for smoking detection involves aligning the accelerometer and gyroscope signals on the same time axis. We use linear interpolation to achieve this timing alignment. We construct the interpolated time-axis to have timesteps when a small amount of data is missing (less than 2 seconds). This allows for filling in missing values in small segments. With the common interpolated time-axis constructed, we interpolate both accelerometer and gyroscope data using linear interpolation. With the accelerometer and gyroscope data aligned to the same axis, next, we apply our activity detection model to remove the segments with high-intensity activity (Walking, Stairs, and Exercise).

5.6.2 Smoking Puff Detection

In our approach to detecting the smoking puff gestures, we build upon the existing works of smoking detection. Our work extends *puffMarker* [10]. Similarly, we aim to detect the smoking puff gesture involving a hand-to-mouth action using inertial wrist sensor data. We elaborate on their methodology before presenting our modifications and novelties.

Building Upon Prior Works By Identifying the Limitations of Existing Puff Detection Model

Figure 5.3 shows four smoking events with multiple puffs in each. Figures 5.3a and 5.3c are from left wrist and Figures 5.3b and 5.3d show smoking events from right wrist. In each sub-figure, we plot the gyroscope magnitude and accelerometer x,y, and z-axis from top to bottom. The straight black lines indicate the labeled segment. The individual figures show that smoking puff gestures are sandwiched between two

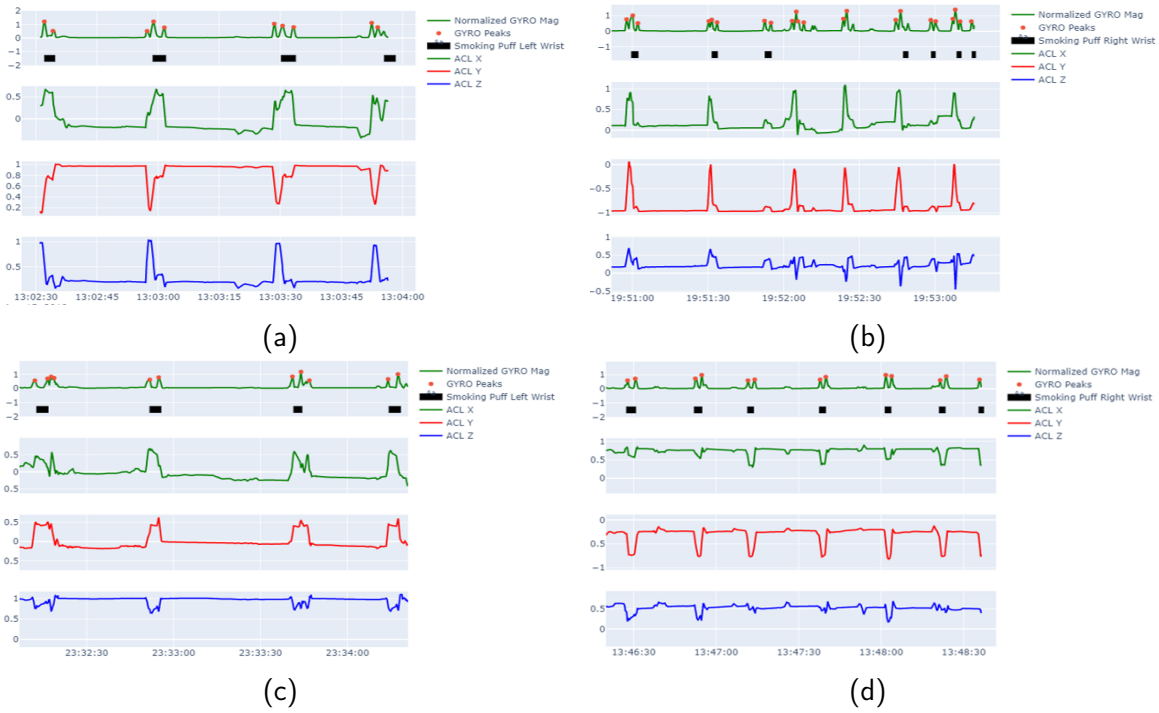


Fig. 5.3: Figures showing inertial wrist sensor data of both wrists during smoking episodes

gyroscope magnitude time series peaks. The first peak indicates the movement of the hand from the body to the mouth. The plateau following the first peak is when the smoke is inhaled. Following this inhalation, the second peak indicates the movement of the hand from the mouth back to the resting position in the body. Therefore, we can trace the start and end of a smoking puff using the gyroscope magnitude data. *puffMarker* uses this approach to identify candidate puffs. Once the candidate segments have been generated, *puffMarker* applies multiple decision rules to accept or discard the candidate segment. The first rule involves checking if the accelerometer y-axis (lateral axis in *puffMarker*) signal moves from low to high (for right hand) and high to low (for left wrist). Without the wrist-specific change, *puffMarker* rejects the candidate segment from further analysis. However, Figure 5.3 shows that the mentioned change in the lateral axis (accelerometer x-axis in our study) does not always manifest in smoking data collected from the field. In the two left wrist smoking events (Figure 5.3a and 5.3c), we observe a low to high transition of lateral axis in each smoking puff. This is contrary to

puffMarker assumption of high to low transition in the case of the left wrist. Also, we observe a low-to-high transition for the right wrist in Figure 5.3b and a high-to-low transition in Figure 5.3d. Thus, *puffMarker* will fail to detect 3 of the 4 smoking events shown in Figure 5.3. The next notable rule involves determining the hand orientation given by the roll and pitch angles. *puffMarker* ensures consistency in roll and pitch calculation by fixing the position and orientation of the wrist sensor in both hands. We discuss the roll and pitch calculation limitations in Section 5.5.4. Our findings show that determining the precise direction of the accelerometer sensor axis aligned to the perpendicular axis is difficult. Hence, thresholding based on pitch angles is ineffective in natural field settings. We extend the works in the literature by carefully considering all the factors which introduce failure modes for puff detection models in the natural field environment. We now describe our candidate segment and smoking puff modeling approach.

Candidate Segment Generation and Selection

In our candidate puff generation approach, we carefully adopt methods that are robust to conditions in the field environment. We do not employ methods vulnerable to unaccounted-for changes in the training data. The significant difference in our methodology lies in the orientation-invariance approach of incorporating different types of sensor placement possible in the field.

We generate candidate segments by detecting smooth gyroscope magnitude time series peaks. Before applying the peak detection algorithm, we first smooth the gyroscope magnitude signal using a moving average window of 0.86 seconds. The smoothing window is selected as 5^{th} percentile of the duration of all labeled smoking puffs in our training data. We choose the minimum peak height for peak detection as the 5^{th} percentile of the gyroscope peaks that fall within or nearby the labeled puff segments. The segment between two consecutive peaks is first run through two different filters based on the duration and intensity of motion. First, we see if the segment has a

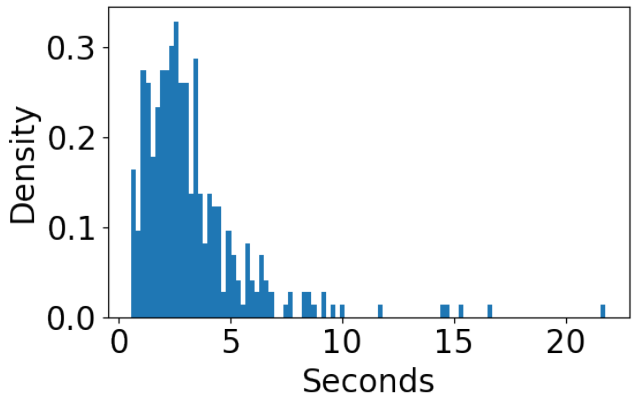


Fig. 5.4: Histogram of the Duration of All Labelled Smoking Puffs

duration within a specific range. We exclude the segment if the segment's duration is lower than 0.86 seconds or longer than 7.43 seconds (95^{th} percentile of the duration of all labeled puffs). The second filter is related to the degree of motion of our wrist within the candidate puff segment. We measure motion using the standard deviation of the accelerometer magnitude time series. If the standard deviation of the candidate segment is above the 95^{th} percentile of the standard deviations of all labeled puffs, we exclude it. The steps mentioned so far are standard processing steps followed in [10]. In the final step, we check for a transition in the accelerometer lateral axis within the puff segment. Unlike [10], we do not impose strict wrist-specific transition criteria. We consider the segment a candidate puff if a low-to-high or high-to-low transition is present.

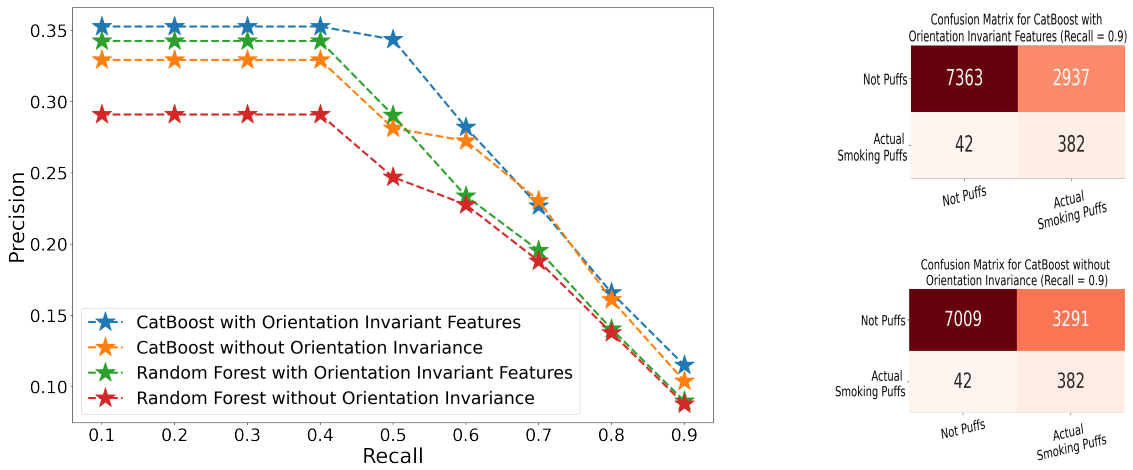
Orientation Invariant Features from Candidate Puffs

We compute several features from the candidate puff segment to compute the feature vector before training our smoking puff detection model. We first increase the width to account for the surroundings of the segment at hand. If (t_1, t_2) is the generated candidate segment, we use $(t_1 - \delta, t_2 - \delta)$ as the whole segment. We select $\delta = 1s$ since it does not overwhelm the original segment. We widened the segment to account for the movement of the hand before reaching the peak value. This movement is part of the smoking puff gesture and should be accounted for in our analysis. Before

computing features, we also align the accelerometer axis corresponding to the lateral axis towards the direction of the earth's gravity. We follow our proposed methodology from Section 5.5.4 to identify and align the accelerometer axis corresponding to the wrist coordinate system (l, p, v). Next, we compute the roll time series using the accelerometer signals. Roll is defined as the rotation around the lateral axis and is computed using the formula $a_l/\sqrt{a_p^2 + a_v^2}$ where (a_l, a_p, a_v) is an accelerometer sample. The formula for the roll calculation is independent of the exact direction of the perpendicular and vertical axes. Since we do not know the exact direction of the accelerometer axes corresponding to perpendicular and vertical axes, we only use the roll values and refrain from using the pitch and yaw angles commonly used in literature. We also use the accelerometer magnitude time series, which is direction independent. From the candidate puff segment, we compute the mean, median, standard deviation, zero crossing rate, 80^{th}, and 20^{th} percentile of the gyroscope magnitude, accelerometer magnitude, roll, and accelerometer lateral axis time series. We also compute peaks in each series and compute three features - the number of peaks per second, mean, and standard deviation of peak amplitudes. Since we do not know the direction of the accelerometer axes corresponding to the perpendicular and vertical axes, we transform each into two orientation-invariant forms. The first is the magnitude, and the second is the magnitude of the consecutive difference. In total, from 8 different time series, we have 72 features, including the actual duration of the segment. Finally, we include a separate binary feature indicating whether it is left or right wrist. Thus, we have 74 features in total from each segment. We use these features robust to the changes in wrist-sensor placement and orientation in the field.

Puff Detection Model

We first label each candidate puff as an actual puff or not. We apply our candidate puff generation method on data of days where video-coded labeled data is available. Thus, our dataset is imbalanced, with more non-puffs dominating the actual



(a) (b)

Fig. 5.5: Performance of Smoking Puff Detection Model

puff counts. We have $10,300$ non-puffs compared to 424 real puffs. The number of candidate-labeled puffs is higher than that of human-labeled puffs (360). This is because, in rare cases, two consecutive smoking puffs may be very close to each other with only one label coded by the human annotator. Using the features and the associated ground truth labels, we train and optimize a gradient-boosted decision tree model to accurately classify the actual puffs. We employ the CatBoost model [223] (short name for Categorical Boosting) using the open-source CatBoost library. CatBoost is a gradient-boosting tree model that has achieved state-of-the-art results in many tasks and requires comparatively little hyperparameter tuning [224]. We use Gradient Boosted trees since we have many features, and exploiting their mutual interaction is the key to successfully training our model. Our choice of the model is also inspired by its state-of-the-art classification results and ease of parameter tuning compared to traditional models such as Support Vector Machines, Random forest models, and others. CatBoost also includes a built-in parameter when dealing with highly imbalanced datasets. We also train and optimize a Balanced Random Forest Classifier [225] with 500 decision trees employing balanced sampling in each tree.

We train the models using leave one subject out cross-validation. To

demonstrate the effectiveness of our orientation-invariance approach, we also train using features computed without lateral axis alignment and transformation of the accelerometer perpendicular and vertical axes. In training our model, we aim to ensure a high recall of the countable actual puffs in our data. To that end, we experiment with selecting the appropriate probability threshold to obtain a specific recall value.

Figure 5.5a shows the recall vs. precision plot of the trained puff detection models. We can see that Orientation Invariant Features improve the performance of puff detection for each model type. For a recall of 0.5, using orientation invariant features, we obtain a precision of 0.343. In comparison, without using the proposed features, we obtain a precision of 0.28 using the same model. Figure 5.5b shows the impact of our methodology in two confusion matrices. For the same number of true positives, we have 350+ false positives without our approach of orientation invariance.

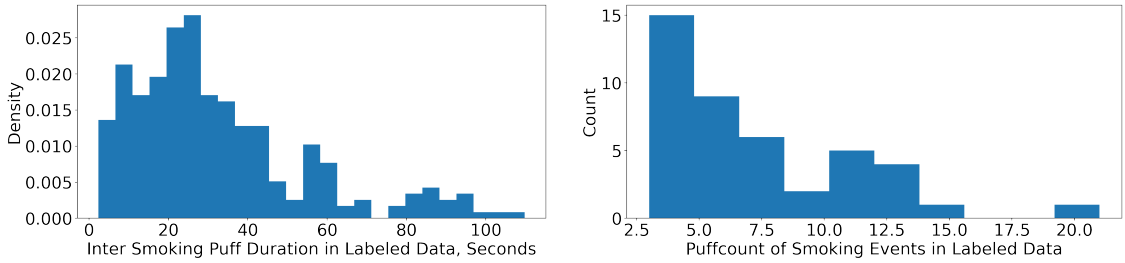
The CatBoost Gradient Boosting model with the proposed feature computation pipeline provides superior performance compared to the Balanced Random Forest Classifier. However, the Random Forest model shows the most significant jump in performance when using the orientation invariance approach.

We select recall of actual puffs as our metric after considering the performance of our model on the training data. Compared to existing works where smoking puff detection models are reported to be highly accurate in field settings and serve as the backbone of the smoking detection process, we find that the performance of actual puff detection is significantly less in our field-collected training dataset. This speaks to the uniqueness of our training data collected in an uncontrolled setting in the field environment. It also illustrates the difficulty in training a smoking detection model for the natural field environment with multiple confounding factors. Hand-to-mouth gestures are common to other activities such as eating, drinking, touching the face, etc. Hence, we must develop a more comprehensive modeling scheme to construct a smoking episode detection approach on top of these noisy detected puffs. We, therefore, propose

our next step of smoking episode construction from noisy detected smoking puffs in the *rSmoke* smoking detection methodology.

5.6.3 Smoking Episode Construction from Noisy Detected Puffs and Event Modelling

We apply the proposed puff detection model to the training data in a leave-one-subject-out cross-validation setting to detect the actual smoking puffs. Using the puff detection model described and with the operating point set at Recall$>= 0.9$, we can filter out 71.4% ($7,363$ out of $10,300$) of the candidate puff segments. However, false positives overwhelm the actual puffs detected (1 to 7.7 true positive to false positive ratio). Existing smoking detection methods construct a smoking episode from detected using a simple rule-based approach of designating a smoking event if a minimum number of puffs are seen within proximity to each other. They employ additional sensor types for effective deployments, such as Respiration [10, 55], smart lighters [61] or quaternions [58]. However, given the nature of our dataset, these methods will prove inadequate using wrist-worn sensors alone. We, therefore, develop a smoking event detection methodology that considers the high rate of false positives and aims to accurately identify smoking events from detected smoking puffs.



(a) (b)

Fig. 5.6: Distribution of the Inter Smoking Puff duration and count in labeled data

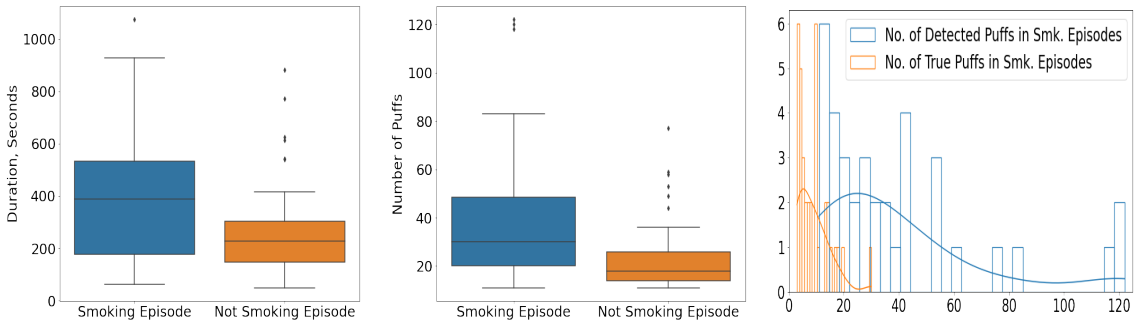
Constructing Candidate Smoking Episodes from Detected Smoking Puffs

In constructing a candidate smoking episode from the detected puffs, we first investigate the presence and patterns of smoking puffs within each episode. We utilize

the distance between consecutive puffs as a measure of starting a new episode. We define a candidate smoking episode as a series of k detected puffs $[(t_i, f_i), (t_{i+1}, f_{i+1}), (t_{i+2}, f_{i+2}), ...(t_{i+k}, f_{i+k})]$ if for $1 <= j <= k$, $t_{i+j} - t_{i+j-1}] \leq \alpha_{duration}$. Imposing an upper bound on the distance between two consecutive detected puffs will isolate spurious detected puffs from being considered part of a smoking episode. We use the domain information from the labeled data to select the appropriate threshold, $\alpha_{duration}$. Figure 5.6a shows the distribution of the time distance between subsequent labeled smoking puffs in our training data. We use the 93^{rd} percentile of the inter-smoking puff duration ($\alpha_{duration} = 78.016$ seconds) as our threshold. The next criterion we employ is the number of detected puffs (k) within a smoking episode. Figure 5.6b shows the distribution of real smoking puffs in individual smoking episodes in the labeled data (mean = 7.07, std = 3.96). We select the minimum of this distribution($k = 3$) as our threshold. Every candidate smoking episode must have at least 3 detected smoking puffs. Next, we designate the candidate episode as an actual smoking episode. We consider a candidate episode an actual one if at least two real smoking puffs are present amongst all the detected puffs within the episode. Using the above methodology, the total number of candidate episodes is 167 with 127 non-smoking episodes and 40 smoking episodes. From now on, we propose our methodology to distinguish between smoking and non-smoking episodes.

Excluding Non-Smoking Episodes Based on Duration and Count of Detected Puffs

In distinguishing between the smoking and non-smoking episodes, we first use rule-based approaches of filtering out obvious non-smoking candidates. Figure 5.7a shows the box plot of the duration of non-smoking and smoking episodes. We observe that the duration of non-smoking episodes is lower when compared to smoking episodes. The 10^{th} percentile of the duration of smoking episodes is 111.645 seconds, equal to the 49^{th} percentile of the duration of non-smoking episodes. We use 111.645 seconds as an



(a) (b) (c)

Fig. 5.7: (a) Distribution of the Duration of Candidate Smoking Episodes, (b) Distribution of the count of actual puffs within candidate smoking episodes, (c) Distribution of detected and smoking puff counts within smoking episodes

episode's minimum duration to be considered a potential smoking one. With the duration filter applied, we check for the number of detected puffs within a smoking and non-smoking episode. Figure 5.7b shows the quantile distribution of the number of detected puffs within smoking and non-smoking episodes. The 10^{th} percentile of detected puffs within a smoking episode is 11.75, equal to the 35^{th} percentile of the number of detected puffs in a non-smoking episode. Hence, we employ a second filter on the minimum number of detected puffs within a candidate episode to be considered for inference. Using the simple rule-based approach, we are left with 107 total candidate episodes with 71 non-smoking and 36 smoking episodes. In the final step, we propose representing these episodes for probabilistic classification using traditional machine-learning models.

Representing Candidate Smoking Episodes for Learning

We must represent the candidate episodes in a manner suitable for classification with maximum exploitation of the embedded information. The major challenge is the few episodes to learn a viable model. Hence, we devise a novel strategy of independently considering portions of each episode before combining them again to make the final prediction. Considering $E_j = [p_1, p_2, p_3, ..., p_k]$ as a sequence of k puffs with y_j as the ground truth label (smoking or non-smoking episode). We find all contiguous

sub-sequence E_j^i such that $l_{min} \leq |E_j^i| \leq l_{max}$ and $|E_j^i| \leq |E_j|$. Here l_{max} and l_{min} are constants denoting the maximum and minimum puff counts. Assume there is n such sub-sequence present for the episode, $E_j = \bigcup_{i=1}^{n} E_j^i$. For $\forall i \forall j$, we consider all sub-episodes E_j^i independently and train a classification model ϕ such that $\phi(E_j^i) = y_j$.

Selecting l_{min} and l_{max}: We create all possible contiguous sub-episodes from the original candidate episode with puff-counts within the range (l_{min}, l_{max}). Each such sub-episode will represent a portion of the original candidate episode. Figure 5.7c show the distributions of the detected puff counts of smoking episodes. The figure also shows the distribution of smoking puff counts within smoking episodes. The mean number of smoking puff counts is 9.03, while the mean number of detected puffs is 40.14. This indicates that 1 actual smoking puff is present in every 4 detected puffs. We choose $(l_{min} = 8$ to have the maximum chance of including at least two consecutive smoking puffs within each sub-episode. The choice of l_{max} is less sensitive since we consider all possible sub-episodes of length $\leq l_{max}$ and $\geq l_{max}$. We choose the value of $l_{max} = 16$. Thus, our methodology breaks down each candidate episode into multiple contiguous sub-episodes with 2 to 4 smoking puffs present.

Features from candidate sub-episodes: We compute features from each sub-episode before training the episode classification model. We first consider the difference in time between consecutive puffs and calculate the standard deviation, 15^{th} and 85^{th} percentile of the time between successive puffs. We also have a feature denoting how many successive puffs overlap. Next, we have the total duration, puff count, and duration per puff count of the whole episode. In each detected puff, we have attributes such as puff probability as given by the smoking puff detection model, mean and standard deviation of the accelerometer, and gyroscope magnitude. We compute the 15^{th} and 85^{th} percentile of these attributes within the episode. Finally, we aim to locate the small number of actual puffs present amid many detected puffs. Hence, we propose identifying the presence of these smoking puffs by featurizing the time difference

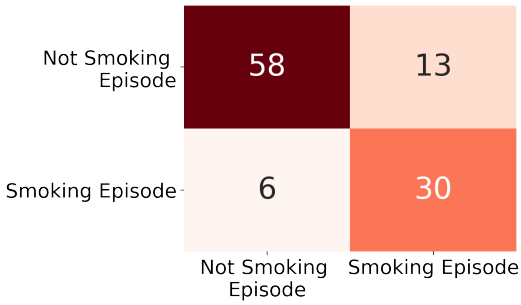


Fig. 5.8: Confusion Matrix for Smoking Event Detection

between all possible detected puffs in the episode. If $E_q^p = [p_1, p_2, p_3, ..., p_r]$ is the sub episode with r detected puffs, $\forall i, i < r-1$, $\forall j, i < j \leq r$, we calculate the difference in time between p_i and p_j and denote it as $t_{i,j}$. Next, we compute features representing the percentage of $t_{i,j}$s fall within the range of specific time duration - $(0, 5), (5, 10), (10, 20), (20, 30)$. The assumption behind this feature is that actual smoking puffs, if present, will be evenly distributed within the episode, and the consistent time difference between them will create mass in specific duration ranges.

Smoking Event Detection Model:

With features from each sub-episode computed, we train a classification model to independently predict the label of each sub-episode given the features from it. We train a CatBoost [223] gradient boosting model in a leave one subject out cross-validation setting. Using this model, we have a probability of being a smoking episode for each sub-episode. The 95^{th} percentile of the probabilities of all sub-episodes belonging to an original episode is denoted as the final probability of being a smoking episode.

Figure 5.8 shows the confusion matrix of smoking episode detection. We obtain a leave one subject out micro f1 score of 0.76, rising to 0.83 for a weighted average, and a roc-auc score of 0.85. In the following section, We further report the performance of our trained model using testing data from the two field studies.

5.7 Performance on Detecting EMA-Reported Smoking Events

We evaluate the performance of *rSmoke* using data from two field studies. Both were smoking cessation research studies involving participants who were regular smokers but wanted to quit. The study protocol involved a pre-quit period when participants went about their daily lives. They quit smoking on their quit day, intending to give up smoking altogether. The post-quit period follows when study coordinators monitor the participants' smoking abstinence status. All the participants wore wrist and chest sensor suites in their natural environment for the whole length of the study. Using EMAs, participants could self-report if they have smoked or not in the recent past. We report the performance of our developed *rSmoke* models to detect the actual occurrence of self-reported smoking events through EMAs. We first explain the details of EMA-based reporting of smoking occurrence and how we construct our ground truth using the self-reports. Then we report the performance of *rSmoke* model in detecting these smoking events.

5.7.1 Self-Reporting Smoking Occurrence using EMAs

Participants filled out regular and random EMAs every day during the study period. The EMAs contained questions about the timings of smoking in their recent past. In the EMAs, the participants were asked, 'Since the last assessment, have you smoked any cigarettes?'. If they responded with 'yes,' we asked them, 'how many cigarettes did you smoke since the last assessment?'. The participants could report the exact number of cigarettes. If they say only 1 cigarette, they are asked, 'How long ago did you smoke the cigarette?'. If they report more than one cigarette, we ask, 'How long ago you smoked your first cig?' and 'Most Recent cig, how long ago?'. The questions were designed to determine the time of smoking and the number of cigarettes they smoked. For reporting the timing of smoking occurrences, participants only indicate a 2-hour time window - '0 - 2 hours ago', '2 - 4 hours ago' and likewise (See Figure 3.7a in Chapter 3 for more details on the EMA reporting questions and answers).

Table 5.1: Performance of *rSmoke* model from smoking self-reports

Test Study Remarks	Number of 2-hour blocks with EMA reported Smoking	Number ' of Participant Days with No Smoking	Model	True Positive Per Day	False Positive Per Day	Precision	Percentage of EMA reported 2 -hour Blocks with smoking events
Original Sensor Mount	804	451	***puffMarker***	1.51	0.62	**0.71**	30%
			rSmoke	1.91	1.01	0.65	**39.4%**
Switched Sensor Mount	373	173	***puffMarker***	0.81	0.52	0.61	12%
			rSmoke	**2.06**	1.07	**0.66**	**34%**

We use these EMAs to design our ground truth labels for smoking occurrences. For true positive analysis, we consider the EMAs where the participants reported smoking events using the 2-hour windows. If $(t, t+2)$ is one 2-hour block where the participant said he smoked cigarettes, we check to see if our model detected smoking within this time range. If a smoking event is present, we consider it a true positive. We report the number of true positives per participant day. We also report the percentage of blocks the participants reported to have smoked contains *rSmoke* detected smoking events.

We consider the days when the participants reported no smoking occurrences for false-positive analysis. On those days, we check their EMA response to see when they responded 'No' to the question, 'Since the last assessment, have you smoked any cigarettes?'. We checked the last assessment time if the participants responded to not smoking. Let (t_1, t_2) be the time window where the participant reported not smoking cigarettes. We counted the number of *rSmoke* detected smoking events in this time window. According to the EMA reports, each smoking event within this window is a false positive. We count the false positives per participant day and report them in our results.

For comparison, we also applied *puffMarker* [10] to both studies. We adopted the original puffmarker using both the wrist-worn IMUs and chest-worn respiration sensors and compare its performance against the wrist-based *rSmoke* model alone.

5.7.2 Results

Table 5.1 shows the performance of the *rSmoke* and *puffMarker* models in the two test studies with original and switched sensor mounts. The *puffMarker* model

employs both the wrist-worn inertial and chest-based respiration sensors. In contrast, *rSmoke* uses the inertial wrist sensors only. We report the number of true and false positives daily in the studies. We also report the percentage of EMA-collected self-reports with the corresponding smoking events detected. We need the precise moments of smoking to train the risk estimation models. Using the sensor-based detection of smoking events and confirmation through EMAs, we can confidently capture the smoking lapse moments. Hence, the percentage of EMA-collected self-reports with the corresponding smoking events plays a vital role in deciding between the detection models. The model with a better correspondence rate will be more suitable for use in the risk estimation model training and deployment.

In the test study with the original sensor mount, *puffMarker* gives a true vs. false positive per day of 1.51 vs. 0.62 while *rSmoke* obtains a true vs. false positive per day of 1.91 vs. 1.01. The improved performance of the *puffMarker* model (71% vs. 65% true positive rate) can be attributed to the usage of respiration sensors in discarding many falsely detected puffs. Since *puffMarker* assumes a fixed orientation of the sensor axes, the number of smoking events detected is low compared to the wrist-only *rSmoke* model. We observe this in the percentage of correspondence with self-reports of smoking events. Using only the wrist-worn inertial sensors, *rSmoke* model obtains a 39.4% self-report correspondece compared to 30% for *puffMarker*.

The major gain in performance is observed in the testing study with switched sensor mounts. *rSmoke* model achieves a true positive per day of 2.06 with a false positive rate of 0.34 compared to 0.81 for *puffMarker* with a false positive rate of 0.39. Despite the use of Respiration sensors in *puffMarker*, the change in inertial sensor mount and the lack of robustness of the methods contribute to an EMA correspondence rate of only 12%. In comparison, *rSmoke* achieves a correspondence rate of 34% using wrist-worn sensors alone.

Overall, the proposed wrist-only *rSmoke* model outperforms the *puffMarker*

model using both the wrist-worn IMUs and chest-based respiration sensors. Also, *rSmoke* being a probabilistic machine learning model allows us to tune the operating point of the smoking-event classifier which can be very useful our use case of smoking lapse risk estimation.

5.8 Limitations and Future Works

This work improved on [10] using only 6-axis wrist-worn inertial sensors for detecting smoking events. However, this work has several limitations, which provide opportunities and challenges for future research. First, there was a paucity of enough training data (only 41 labeled smoking events) from 10 participants. Future works with more training data, especially with more labeled smoking events, can increase the accuracy of puff detection and improve smoking event detection in the field.

Second, we used a simple threshold-based approach to determine whether the sensor is worn. Determining sensor is worn or not using an accelerometer is tricky as it may lead to many false positives. Moreover, gyroscope sensors are susceptible to motion, leading to the generation of many spurious puffs. Future works can improve the method to detect sensor wearing more accurately to limit false positives.

In all three studies, participants wore sensors on both wrists, considered independent in our model. However, in real life, we expect an individual to wear a sensor on either wrist. Hence the smoking detection model needs to consider both wrists together so that the model works for a single sensor on either wrist.

In the field, wrist sensors can be worn with varying degrees of tightness, which can contribute to the noise component in the data. To reduce the noise caused by the tightness of the wearing of the sensors, future works may need to include PPG sensing to filter out segments where the sensor was worn loosely.

Collecting ground truth for smoking behavior is challenging. EMAs provide a secondary evaluation only. The future study design can consider this phenomenon to improve the collection of reliable ground truths.

A personalized model of smoking detection would be more useful in the long term. Online learning methods to calibrate e smoking model for each participant can be done in the future.

This work presented cigarette smoking event detection using wrist inertial sensors. Future works can also expand this work to vaping, e-cigarette smoking, and others. Further works can also use GPS sensors to confirm smoking instances and reduce false positive samples.

5.9 Chapter Summary

Nearly one of every five deaths in the United States is caused by smoking [226]. Automated detection of smoking from wrist sensors can facilitate the precise intervention or delivery of the treatment at the right moment for an abstinent smoker to prevent relapsing. Variability in sensor configurations, sensor placement resulting in variability in axes orientation, lack of sufficient training data, and difficulty in collecting reliable ground truths present challenges in building a robust smoking detection model from wrist sensors in the field. Difficulty distinguishing smoking from similar behaviors involving hand-to-mouth gestures such as eating, brushing, or drinking and determining whether the sensor is worn can result in many false positives. This work provides an orientation-invariant modeling framework for detecting smoking events from wrist-worn inertial sensors. The model achieves a precision of at least 0.65% two different test studies with higher self-report recall rate than existing models. This work could be leveraged by researchers or health practitioners to automatically detect smoking events in the field to assist in intervention and treatment decisions to prevent any abstinent smokers from relapsing.

Chapter 6

Continuous Assessment of Smoking Lapse Risk From Wrist-worn Sensors

6.1 Introduction

In this chapter, we combine all the completed works from previous chapters to realize our goal of developing a continuous smoking lapse risk assessment model using data collected from wrist-worn sensors alone. In Chapter 3, we developed *mRisk*, a continuous lapse risk estimation model in the participants' natural environment. We developed *mRisk* using data from chest-worn sensors. The different modalities of sensor data were first transformed into an intermediate representation of dynamic risk factors using state-of-the-art machine learning models. From chest-worn ECG and Respiration sensors, we computed the continuous stress levels of participants in their natural environment. From chest-worn accelerometer sensors, we computed the physical activity level. And finally, using the GPS-based location information collected through smartphones, we computed the proximity to personal and general smoking spots. Using this psychological (stress), behavioral (activity), and contextual (proximity to smoking spots, location) information, we trained a deep neural network-based model to continuously output the impending risk of smoking lapse. We evaluated the *mRisk* models based on their ability to inform the delivery of just-in-time smoking interventions to avert the occurrence of lapse in the near future. The findings from *mRisk* show that continuous lapse risk estimation from chest-worn mobile health sensor data is feasible. However, accomplishing the overall goal of assessing smoking lapse risk from wrist-worn sensors involved solving multiple challenges facing human psychological and behavioral state prediction using machine learning models.

The first challenge concerns estimating stress and activity levels from wrist-worn PPG sensors. In Chapter 4, we proposed methods to enable robust inference of stress and activity from wrist-worn sensor data. We developed *CQP*, a continuous stress assessment model from PPG sensor data that proposes to integrate the raw signal

quality of PPG signals as a tool for selecting viable sensor segments and accurate computation of heartbeat-related features. We trained a convolutional neural network (CNN) based physical activity inference model with publicly available labeled data to estimate activity levels. We train the activity detection model to output physical activity labels from the magnitude of the accelerometer signal in fixed-length windows.

The next challenge stems from developing a smoking detection model from wrist-worn inertial sensors for deployment in the natural field environment. Current smoking detection models' limitations include a rigid approach of assuming a fixed configuration of inertial sensor axes and a lack of robustness against sensor orientation and placement changes. We developed *rSmoke* in Chapter 5, an orientation-invariant approach to detecting smoking events from accelerometer and gyroscope sensors in the field. Our proposed *rSmoke* model increases recall of smoking self-reports, thus providing us with improved data coverage for training our smoking risk estimation model.

With wrist-based inference of lapse risk factors and smoking events in the natural environment proven possible, we move towards training and developing the *mRisk* models from multi-modal wrist sensors in participants' natural environment. We propose an end-to-end methodology for estimating smoking lapse risk using wrist-worn PPG, accelerometry, and GPS sensor data. Our approach considers the intricacies of processing wrist-sensor-derived inference of dynamic risk factors in the wild for use as model inputs. Like the *mRisk* models developed using chest sensor data, we evaluate our models based on their ability to inform intelligent design and delivery of just-in-time smoking interventions in the wild. Previously in Chapter 3, we simulated a threshold-based design of the intervention delivery mechanism where we deliver an intervention whenever the risk score is higher than a threshold and sufficient time has passed since the previous intervention. We propose a more useful intervention delivery scheme that simulates the delivery of interventions at the onset of a risk episode. In developing our methods, we consider various challenges concerning the triggering of

smoking interventions to participants in their natural environment. We need to deliver the intervention to be effective for participants. The delivery mechanism must be simple enough for implementation on a mobile device. Also, the frequency of interventions must be tolerable to avoid fatigue [172].

We first describe the smoking cessation dataset where participants wore both the chest-worn ECG, Respiration sensors alongside wrist-worn PPG and inertial motion sensor units. Second, we provide an overview of the specific data processing routines unique to the wrist sensor data for estimating smoking lapse risk. We train the *mRisk* models using wrist-sensor-based inferences of dynamic risk factors representing participants' psychological, behavioral, and environmental contexts. Next, we describe our methodology for delivering effective interventions to participants in their natural environment. Finally, we present our results of simulated intervention delivery using risk scores produced by the trained models. We compare and contrast the results obtained using wrist and chest sensor-derived inferences of dynamic risk factors.

6.2 Smoking Cessation Research Study with both Wrist and Chest Sensors

We describe the smoking cessation research study involving the use of both chest and wrist sensors. This is the same study from Chapter 4. The Institutional Review Board (IRB) approved the study and all participants provided written consent.

6.2.1 Study Participants Recruitment and Protocol

The participants had to be regular smokers for the last two years and willing to quit to be eligible to take part in the study. Recruitment flyers were posted in public areas such as college campuses, community clinics, churches, and local restaurants and bars to recruit participants. Advertisements were placed in local newspapers and on the radio. In-person recruitment was implemented to promote enrollment when requested by groups or institutions with a population that is likely eligible and interested (similar to the smoking cessation research study in Chapter 3). The recruited participants went through the informed consent process during their initial (baseline) lab visit.

In their visit to the lab, participants were trained in the proper use of the sensor devices and how to respond to questionnaires in the form of Ecological Momentary Assessments (EMA) via a study-provided smartphone. They wore the sensors for 4 days during the *pre-quit* phase. On their preset quit date, participants returned to the lab. Then they wore the sensors for 10 more days during the *post-quit* (or *smoking cessation*) phase. The participants were compensated for completing in-person visits — $30 each for Visits 1, 2, and 3, $80 for Visit 4, and $60 for Visit 5. They were compensated at the rate of $1.25 for completing each smartphone survey if they wore the on-body sensors and/or collected usable sensor data at least 60% of the time since the last phone survey, and $0.50, otherwise for completing each smartphone survey. The participants were also reimbursed for parking or bus tokens to defray the cost of traveling to the project site.

6.2.2 Wearable Sensor Suites

Participants wore both the chest and wrist sensor suites in their daily lives. We employed the AutoSense [72] chest sensor suite, which contains ECG (64 Hz), Respiration (21.33 Hz), and Accelerometry (16.33 Hz) sensors in a chest belt. Participants also wore wristbands fitted with PPG, Accelerometer, and Gyroscope sensors in both wrists. The sampling frequency of all the sensors in the wristband is 25 Hz. Participants were also given a study smartphone to complete EMAs and required questionnaires. The smartphone comes with the open-source mCerebrum software [156] installed. The smartphone also collects GPS data. However, we do not constantly sample GPS sensor data to conserve battery energy. We only sample GPS location whenever smartphone inertial sensors register movement. We employ this strategy for optimizing battery usage and collecting data for longer periods between charging.

6.2.3 Data Volume

Out of the 110 participants who completed the study, we have sufficient data from 54 participants from their pre-quit period. From these 54 participants, we have data from 568 participant days. From the wrists, we have $7,056$ hours of stress data

(12.42 hours per day) and $9,979$ hours of activity data (17.56 hours per day). From the chest, we have $7,353$ hours of activity data (12.94 hours per day) and $4,892$ hours of stress data (8.61 hours per day). Additionally, we have $9,366$ hours of location data (16.49 hours per day). The 54 participants completed a total of $1,388$ EMAs. We obtain the smoking lapse times from these 54 users using EMA-collected self-reports and the *rSmoke* model from Chapter 5. Although all 54 participants lapsed according to their EMA response, we only have confirmed lapse events for 35 of them. From the 35 participants, we have 98 confirmed lapse events in total.

6.3 Wrist-based Smoking Lapse Risk Estimation

We aim to train and develop the *mRisk* models from Chapter 3 using data derived from only wrist-worn sensors. This section describes the data processing routines for transforming the wrist-worn sensor data into intermediate representations of smoking lapse risk factors. We touch on the feature computation and processing pipelines specific to wrist data. Next, using the features, we briefly describe the training and evaluation methodology for the lapse risk models.

6.3.1 Data Processing

We use the stress and activity inference models from Chapter 4. We apply the models on our dataset to compute the continuous stress from PPG and physical activity labels from wrist-worn accelerometer sensors. The stress model outputs a continuous time series representing momentary stress levels. We compute stress episodes from the stress likelihood time series. Episodes indicate the locations where participants were subject to major stress cues. Episodes help us capture the historical cues of stress influence. We call these episodes 'stress events-of-influence.'

The activity detection model outputs one of five discrete activity labels every 20 seconds - Stationary, Walking, Exercise, Stairs, and Sports. We consider Exercise, Walking, and Climbing Stairs high-intensity activities compared to Stationary and Sports. We first compute a minute-level binary time series indicating if there is

high-intensity activity present within a minute. This continuous activity time series reflects whether a participant was active. From the activity time series, we compute activity episodes indicating the contiguous duration of activity in periods. We term these as 'activity events-of-influence.' Both stress and activity suffer from spurious missingness owing to a lack of viable sensor data, the sensor not being worn, and other factors. Before computing the events of influence, we impute the small gaps in the stress and activity series.

The GPS sampling scheme in our study follows a battery-saving principle to ensure a longer data collection period. The GPS sensors sample the location data whenever smartphone accelerometer sensors indicate motion. This creates a sparse representation of the participant's location compared to the dense representation produced by GPS sampling every second (in the Smoking Cessation Research Study of Chapter 3). We adopt the methods proposed in [32] to process GPS samples into representations of location contexts.

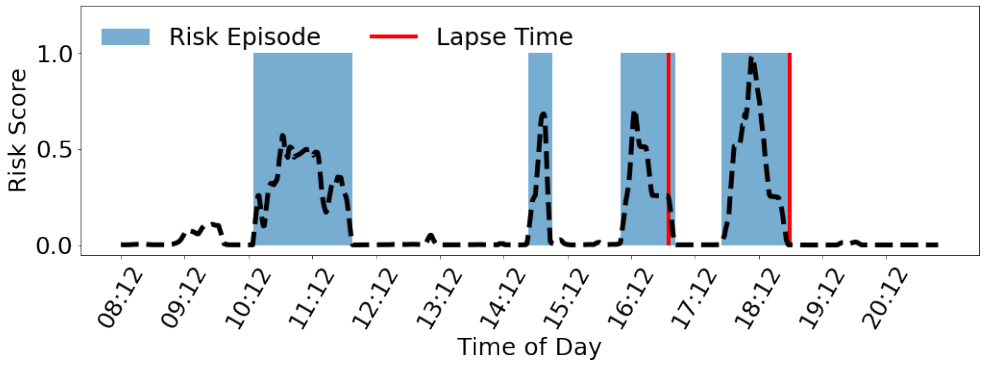
The first step in pre-processing GPS data is employing temporal clustering to sparsify the original time series into stay times at dwell places. For example (t_1, t_2, d_1) represent the participant stayed at dwell place d_1 from time t_1 to t_2. The stay time clustering is necessary for reducing the number of GPS samples before employing computationally expensive density-based spatial clustering methods [227, 228]. Since we employ conditional sampling of the GPS sensor in the first place, the number of GPS samples is not very high. We can apply density-based clustering in the first step. We first de-noise the GPS time series by removing outlier samples with low accuracy. Next, we apply a density-based clustering method to derive the participant dwell places in their pre-quit period. A dwell place designates a location where the participant stayed for a while. Based on the duration of stay and frequency of visits, we mark each dwell place as Significant or Transient. The significant dwell places are specific to participants, are comparatively small in number, and represent locations such as homes or offices where

the participant resided for a significant amount of time. In contrast, transient places represent occasional visits to places such as shops, places of worship, hospitals, etc. The number of transient places per participant is also quite large.

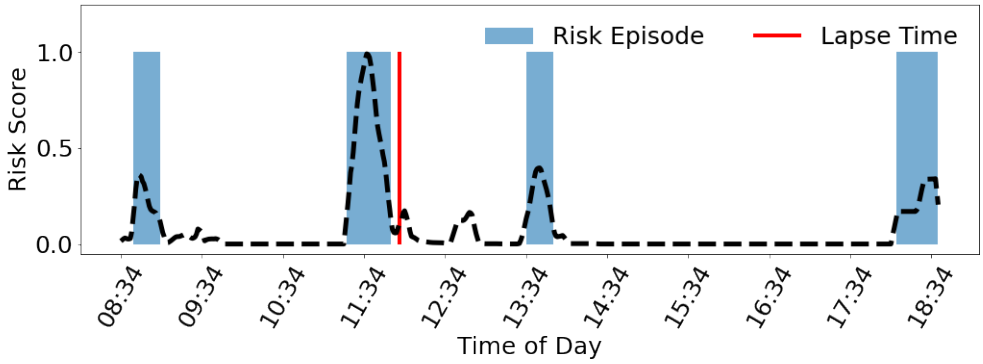
Using the exact timings of smoking events given by the *rSmoke* model and confirmed through EMAs, we can designate all the dwell places as personal or general smoking spots. Personal smoking spots are significant dwell places where participants smoke regularly. We compute the general smoking spots from the transient dwell places. We extract the semantic meaning of these transient places to assess if it is potentially conducive to smoking. Using a customized Point-of-Interest (POI) database, we label each transient place as having one of 6 semantic types - alcohol, cigarette point-of-sale, retail, medical, school, and places of worship. Using smoking allowance level (provided by EMA), semantic meaning, and occasional observed smoking events, we designate some transient places as general smoking spots. These spots indicate public locations conducive to smoking and can illicit a smoking lapse behavior in a visiting participant. Once we obtain the locations of significant and transient dwell places along with the smoking spot classifications from the pre-quit period, we can represent the post-quit GPS data of all participants into representations of smoking spot visitations and proximity to smoking spots.

6.3.2 Risk Estimation Models

Using the stress, activity, and location contexts derived from wrist-sensor data, we train the LSTM-based deep neural network *mRisk* models from Chapter 3. The first model is called *DRES*, which stands for Deep Model with Recent Event Summarization. The second model is called *DDHI* - Deep Model with Historical Influence. Both models utilize continuous inference features to represent the current physiological, behavioral, and environmental context. We use two approaches to represent historical context as given by the events-of-influence time series (stress, activity, and visits to smoking spots). The *DRES* model uses innovative features to represent the events-of-influence time



(a)



(b)

Fig. 6.1: Examples of Lapse Risk Episodes in daily Smoking Lapse Risk Scores

series. In *DDHI*, we forego feature engineering and employ a novel heterogeneous event-encoding methodology first proposed in Chapter 3.

We also employ the novel Rare Positive Loss function proposed in Chapter 3 to train these models using data derived from wrist-worn sensors. We train the models in leave-one-subject-out cross-validation and derive the continuous lapse risk scores. The risk scores given by the trained models are evaluated based on their ability to inform just-in-time smoking interventions. In this chapter, we take it further and introduce the concept of risk episodes. These episodes are useful in determining the appropriate strategy for designing an effective smoking intervention methodology.

6.4 Intervention Delivery Informed by Risk Episodes

Figure 6.1 shows daily plots of smoking lapse risk given by the *DDHI* model. We train the model using wrist-sensor-based stress, activity inferences, and location contexts derived from smartphone-collected GPS samples. We overlay the plots with vertical lines showing smoking lapse times. The goal is to enable intervention delivery at opportune moments to maximize our chances of averting impending lapse moments. Our models output a continuous risk score every minute. We need to make an informed decision at every minute about whether we want to send the intervention. In Chapter 3, we simulate the intervention delivery mechanism to deliver interventions whenever the risk score crosses a certain threshold and a minimum time has passed from the previous intervention time. The methodology is oblivious to the local characteristics of the smoking lapse risk time series around the intervention point. A spurious high-risk score value can trigger an intervention with a high chance that the intervention does not address the actual level of lapse risk the participant currently experiences. In this section, we first introduce the concept of risk episodes from continuous lapse risk scores. The episodes indicate times when the participant experienced a rise in risk levels, followed by the risk levels again falling back to the baseline. Figure 6.1a shows an example of four risk episodes the participant experiences in a single day. Two episodes precede ground truth lapse moments confirmed by EMA-collected self-reports. Figure 6.1b also shows similar phenomena with a risk episode preceding a smoking lapse behavior. The episodes, if properly identified, provide the opportunity to construct proactive intervention delivery mechanisms.

6.4.1 Finding Opportune Moments for Smoking Interventions

We base our proposed intervention delivery method on utilizing the episodes to deliver effective interventions. Our method of intervention delivery also needs to be simple enough to be implemented in mobile devices without the need for complex processing routines. Most of all, the proposed method must be fit for online

implementation. Detecting the whole episode and then delivering intervention based on the whole will delay the delivery time of interventions. This may render the intervention ineffective, with high chances of participants lapsing beforehand. Hence, we do not go into the details of constructing risk episodes in their entirety. We propose methods to directly identify the opportune moments of intervention delivery within a risk episode without constructing the whole episode.

We observe in the example figures that each risk episode starts with a rise from its baseline and reaches a peak risk value. This peak risk time indicates the participant's highest risk level. We propose to detect these peaks in an online manner. As the model outputs continuous risk scores, we check whether we are at the peak risk value. If y_i is a peak, then $y_i > y_{i-1}$ and $y_i > y_{i+1}$. This is a simple first-order condition. And in the presence of noise, we can be overwhelmed with spurious peaks in the risk time series. To reduce the noise, we propose transforming the raw risk score time series by applying an online moving average of fixed samples. We then proceed to compute peaks in the average risk score time series. To be stringent in selecting peaks, we also introduce a second-order condition. If y_i is a peak, then $y_i > y_{i-1}$, $y_{i-1} > y_{i-2}$, $y_i > y_{i+1}$ and $y_{i+1} > y_{i+2}$. Notice with the second-order condition, we will only find the episode peak 2 minutes after it occurs. Next, we impose two more criteria on peak selection. The first criterion is the peak amplitude. The value of the risk at the peak point has to be greater than a threshold θ. The second criterion concerns the rate of rise in a risk episode from the start to the peak location. Let $y_i > \theta$ indicate a peak location in the risk time series that fulfills the second-order peak condition. Let $y_j, j < i$ be the valley that indicates the start of the episode to which y_i belongs. Our second criterion is based on the amount of risk from the start point j to the peak location i. We only propose to deliver an intervention at time $i + 2$ if $\sum_{k=j}^{i} y_k \geq \delta$. The value of parameters θ and δ will determine the overall frequency of smoking lapse interventions to be delivered.

Table 6.1: Intervention Hit Rate at different daily frequencies of intervention using wrist sensors

		Intervention Frequency per day											
Model	Intervention Delivery Algorithm	3.00	3.50	4.00	4.50	5.00	5.50	6.00	6.50	7.00	7.50	8.00	Median IHR
Random	-	0.31	0.34	0.42	0.46	0.48	0.55	0.61	0.64	0.70	0.76	0.76	0.55
DRES	Threshold Based	0.40	0.42	0.51	0.57	0.58	0.67	0.76	0.83	0.86	0.90	0.91	0.67
	Risk Peak Detection	0.53	0.60	0.62	0.63	0.65	0.71	0.76	0.76	0.77	0.79	0.82	0.71
DDHI	Threshold Based	0.51	0.54	0.63	0.68	0.70	0.76	0.84	0.86	0.89	0.92	0.94	0.76
	Risk Peak Detection	0.55	0.64	0.68	0.70	0.73	0.77	0.81	0.83	0.86	0.86	0.86	0.77

6.5 Evaluating the Effectiveness of Risk Peak Triggered Simulated Interventions

To understand the effectiveness of the trained lapse risk prediction models from wrist sensors, we adopt the approaches from Section 3.7.3, Chapter 3 to simulate intervention delivery at opportune moments indicated by the risk scores. We report the *Intervention Hit Rate (IHR)* values for a given model and intervention delivery algorithm. *IHR* denotes the proportion of lapse events occurring within 60 minutes of intervention. We design our intervention delivery mechanism by triggering interventions at times of risk peaks. We vary the values of parameters θ and δ. This translates to varying levels of aggregate intervention frequency in the whole dataset. We compute the *IHR* values for a given median intervention frequency per day. Next, we interpolate them in the range of $[3, 8]$ interventions per day.

We also adopt two methods for baseline comparison. First, we show the *IHR* results obtained by delivering interventions at randomly selected times within selected time blocks. This method may be used when no risk assessment via sensors or self-reports is available. We divide each waking day into k blocks of time and randomly assign an intervention within each block. By varying the value of k, we can tune the median number of interventions given per day. Second, we adopt the threshold-based intervention delivery method proposed in Chapter 3. Interventions are delivered when

Table 6.2: Intervention Hit Rate of the DDHI model at different daily frequencies of intervention using different sensing modalities

		Intervention Frequency per day											
Sensing Modality	Intervention Delivery Algorithm	3.00	3.50	4.00	4.50	5.00	5.50	6.00	6.50	7.00	7.50	8.00	Median IHR
Chest	Threshold Based	0.53	0.59	0.67	0.72	0.74	0.78	0.77	0.86	0.86	0.92	0.96	0.77
	Risk Peak Detection	0.61	0.61	0.61	0.63	0.71	0.79	0.82	0.81	0.86	0.88	0.88	0.79
Wrist	Threshold Based	0.51	0.54	0.63	0.68	0.70	0.76	0.84	0.86	0.89	0.92	0.94	0.76
	Risk Peak Detection	0.55	0.64	0.68	0.70	0.73	0.77	0.81	0.83	0.86	0.86	0.86	0.77

the risk for lapse rises above a pre-determined threshold (T_L) and at least *intervention gap (I_G)* minutes have elapsed since the last intervention.

Table 6.1 shows the *IHR* results for both the *DDHI* and *DRES* models trained using data collected from wrist-worn sensors alone. The smoking lapse risk models outperform the random baseline by a significant margin. The *DDHI* model with our novel event encoding methodology gives better results than the feature-based *DRES* model.

For each model type, we also report the performance of both the threshold based as well as the proposed risk peak-triggered intervention delivery method. Our proposed intervention delivery method provides superior performance, especially at lower frequencies of interventions. At 4 interventions per day, using the *DDHI* model we obtain 0.68 *IHR* value which is 0.05 larger than using the threshold-based method used in Chapter 3. The performance improvement is even larger (0.62 compared to 0.52) for the *DRES model*. The gain in performance at lower frequencies from using the risk episode peaks is because peaks are stable estimates of high-risk situations, and risk episodes with high-risk amplitudes precede most lapses. As we increase the number of intervention frequencies daily, the threshold-based delivery method outperforms the risk peak-based delivery of smoking interventions. The threshold-based intervention delivery

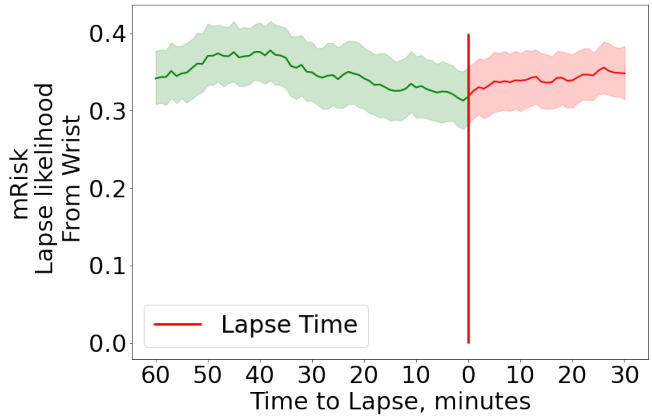


Fig. 6.2: Wrist-derived *mRisk* Lapse likelihood averaged across all the lapse moments

mechanism also has a higher ceiling with 94% of the lapses captured at the cost of 8 interventions per day.

Table 6.2 compares the performance between the *DDHI* models built using the wrist and chest sensors. The wrist-based *DDHI* model performs similarly to the chest-based model. We obtain a median *IHR* value of 0.79 using the chest sensors and 0.77 using wrist sensors only. This demonstrates that our proposed methods of stress, activity, and smoking inferences from wrist-sensor data enable accurate modeling of the smoking lapse phenomenon. The results using multiple sensing modalities also speak to the generalization ability of the modeling approaches proposed in Chapter 3.

We aggregate the risk scores across all lapse moments from all participants to see the general pattern of risk before and after the lapse. Figure 6.2 shows the mean lapse risk (with a confidence interval of 90%) before and after a smoking lapse. The risk score rises and peaks around 50-45 minutes before a smoking lapse. The risk then decreases as the lapse moment approaches. The aggregate rise in risk score before the lapse moment is similar to what we observe in Chapter 3, Figure 3.8b for the chest-based model where we observe a rise in risk 44 minutes before lapse.

Overall, our results show that the proposed wrist-based estimation of smoking

lapse risk is accurate in comparison to the chest-based models and allows the design of effective intervention methods to avert smoking lapse behavior in abstinent participants.

6.6 Discussions, Limitations and Future Works

The proposed methods have several limitations that present exciting opportunities for future research endeavors in computing and health research. The limitations related to smoking lapse risk models include the inability to use temporally imprecise label sources, the lack of ground truth labels of low-risk states, and the presence of only stress, activity, and location features. We explain these limitations in detail in Chapter 3. In this section, we focus on the limitations and future research areas related to wrist-based sensing, smoking lapse risk estimation from wrist sensors, and just-in-time smoking intervention design.

First, we only have viable data from 35 participants only for modeling the smoking lapse risk phenomena using wrist sensors. Increasing the number of participants and data volume will allow our model to learn the diverse interwoven pattern of smoking lapse behavior and will certainly add to the robustness of the learned model.

The second limitation concerns the lack of an online methodology for smoking lapse risk models. Our work does not address the complexity of online estimation of smoking lapse risks in the natural environment. Real-time smoking risk estimation will significantly increase the utility of our models. Future research can work to address this gap by adapting our methods to work with only current and past data. Exploring online learning methods to enable personalized risk estimation using ground truth risk moments from the post-quit period will add to the feasibility of practical deployment. This will allow more accurate estimation of participants' risk contextualized to their environment.

Third, our wrist-based risk estimation model does not consider any fundamental characteristics of mobile health sensing using wrist sensors. Although we use intermediate representations of the raw sensor data with state-of-the-art machine learning models, information about the different sensing mechanisms adopted in wrists

can contribute more information. For example, wrist-worn PPG sensors are susceptible to higher motion artifacts than chest-worn ECG. Our proposed stress model incorporates an independent data quality estimation pipeline to quantify the quality of collected data and aid in stress inference from PPG. In the future, we can consider the quality likelihood time series as an independent information channel for the smoking risk estimation model.

Fourth, we notice the promise of uncertainty quantification of the output risk scores. A measure of uncertainty accompanying the risk allows for more intelligent and rational decision-making on the time and type of interventions to deliver based on the risk score. Uncertainty estimation can account for the uncertainty inherent to the model itself. The risk prediction model is designed to output smoking lapse risk in diverse situations, which may indicate a shift from training data. An estimate of the model uncertainty aims to provide a performance bound and draw the limitations of the lapse risk model [229].

Fourth, we depend on self-reports and automated detection of smoking events to construct ground truth labels of high-risk states. Automated smoking detection models, including the *rSmoke* model presented in Chapter 5, do not have the necessary precision to locate smoking lapse events in the wild without confirmation from self-reports. Thus, we depend on temporally imprecise sources such as EMAs to confirm detected smoking events. Future work can explore novel ways to incorporate information from temporally imprecise sources such as EMAs or incorporate smoking event markings.

Finally, our design of intervention delivery and simulation experiments using risk scores from the trained models is not validated with real-life experiments. We can complete the evaluation of our proposed methods by designing a micro-randomized trial based on intervention delivery informed by smoking lapse risk scores.

6.7 Chapter Summary

In Chapter 3, we employed chest-based ECG and inertial motion sensors for passive sensing of human health and wellness states. Owing to the inconvenience of

wearing chest-wrapped devices daily, the practical utility of using such sensors outside of academic use remains limited. Conversely, wrist-worn wearables or sensor-fitted smartwatches have seen growing adoption in research and commercial use due to their convenient form and the ability for continuous monitoring. Developing a wrist-only smoking lapse risk estimation model from wrist-worn wearable sensors can significantly boost the practical utility and usability. Our work realizes this goal by developing a first-of-its-kind smoking lapse risk estimation model from wrist-worn sensors. We also propose a new intervention delivery mechanism inspired by the episodic characteristics of the risk scores. Simulated results show that our wrist-based smoking lapse risk estimation model can capture 68% of the confirmed lapses at 4 interventions per day. This work opens the door to many exciting research opportunities to increase the rate of smoking abstinence using mobile health sensors.

Chapter 7

Conclusion and Future Directions

7.1 Summary and Key Contributions

In the continuing fight to curb smoking and the tobacco use pandemic, innovative methods relying on passive mobile sensing can usher in a new era of progress and positive results. Precision medicine based on wearable sensing has the potential for delivering just-in-time adaptive interventions to abstinent smokers when they are most vulnerable. Our dissertation adopts this objective and proposes an end-to-end methodology for estimating the imminent risk of smoking lapse behavior. Our method relies on passive and continuous detection of smoking lapse risk factors in the natural environment of individuals to train a novel deep neural network-based smoking lapse risk estimation model from convenient wrist-worn sensors.

Our model relies on several key innovations to address the challenges of continuously estimating smoking lapse risk using wrist-worn mobile health sensors. We propose to represent participants' current and historical context in their natural environment using inferences of lapse risk factors from state-of-the-art machine learning models. We represent the physiological, behavioral, and environmental contexts using sensor-based inferences of stress, activity, and location contexts. We derive these contexts by employing chest-worn wearable sensors and propose an end-to-end smoking lapse risk estimation pipeline. We train two LSTM-based deep neural network models, each with a unique strategy for capturing historical context. We also propose a novel event-encoding methodology of automatically representing the historical contexts captured by events-of-influence time series. To accurately optimize our models with incomplete positive-only labels, we propose a novel loss function. We evaluate the utility of our models based on their ability to inform the design and delivery of just-in-time adaptive smoking interventions. Our preliminary results indicate that using

chest-sensor-derived representations of stress, activity, and location contexts, our model can deliver just-in-time smoking interventions before most of the confirmed lapses.

Our proposed methods of accurate smoking risk estimation from chest-worn wearables open up a new chapter of computing research. Intervening with at-risk participants to avoid impending lapse can improve the overall smoking abstinence rate. However, adopting the chest-sensor-based risk estimation models for widespread use is not straightforward. Outside academic research, the utility of chest sensors is limited due to the inconvenience of wearing chest-wrapped devices daily. Wrist-based sensors, on the other hand, provide a convenient form of wearing and have seen growing adoption in research and commercial use due to their ability for continuous monitoring. Adapting our methodology to enable continuous estimation of smoking lapse risk from wearable wrist sensors is bound to improve the practical utility of our models. Hence, we focus on translating the developed lapse risk estimation methods to work with wrist sensors alone.

To develop the *mRisk* models from wrist sensors, it is imperative to obtain a passive and continuous estimation of dynamic risk factors. To enable robust inference from noisy wrist-worn sensor data, we propose CQP. *CQP* is a machine learning-based data quality indicator, which informs the quality of inference from time-varying signals. We use *CQP* to devise a novel approach of auxiliary estimation and deep integration of signal quality metrics within the inference process. Integrating signal quality levels within the inference mechanism enhances the accuracy and robustness of continuous inference from wrist-worn PPG sensor data compared to existing methods. We also train a deep neural network-based activity detection model to output continuous activity labels using the magnitude of wrist-worn accelerometer sensor data.

To enable wrist-based estimation of smoking lapse risk, we need a working smoking detection model that can accurately detect smoking events from wrist-worn inertial sensor data in the field. We first identify the limitations of existing wrist-based smoking detection methods in the literature. These methods suffer from a lack of

robustness to many possible scenarios in the field related to sensor configurations, orientation changes, and other factors. We propose *rSmoke*, an orientation-invariant approach to first identifying the axes configuration for inertial sensors in the wild. *rSmoke* computes robust features from the inertial wrist-worn IMU data to first detect smoking puffs. To construct smoking episodes from noisy and spurious smoking puffs detected in the field, *rSmoke* proposes a novel smoking episode construction scheme that allows for the representation and identification of smoking episodes. Using the *rSmoke* model, we capture smoking lapse events in the post-quit smoking abstinence period.

We train the proposed *mRisk* models from all the proposed wrist-based inferences to output smoking lapse risk scores from wrist-worn wearables. To simulate the ability of the trained models to deliver intelligent smoking interventions to abstinent participants, we propose a new online intervention delivery mechanism based on risk episodes. Our results indicate that smoking lapse risk scores from the *mRisk* models trained on wrist-based inferences of lapse risk factors.

7.2 Future Research Directions

Our dissertation is the first to propose a comprehensive approach to predict the imminent risk of smoking lapse from wrist-worn wearables. Our work explores many directions to making smoking lapse estimation from wrist sensors a reality. We innovate novel solutions to fill the existing research gaps and propose a detailed end-to-end pipeline of continuous smoking lapse risk estimation from wrist-worn wearables. Our dissertation touches on multiple sub-problems in different fields and employs novel approaches to address each of them. Our proposed methods has limitations that present exciting opportunities for future computing and behavioral research.

In future, the developed risk prediction model can continuously output the real-time risk of a smoking lapse in the natural environment. For effective deployment of the model in a smoking cessation study, future research can address the challenges of building and deploying a real-time online model. Developing an online lapse risk

estimation model to output composite risk scores in real-time will significantly increase the utility of our work. Exploring online learning methods to enable personalized risk estimation using ground truth risk moments from the post-quit period will add to the feasibility of practical deployment. This will allow a more accurate estimation of participants' risk contextualized to their environment.

The presence of many external and internal factors in the natural environment can adversely impact the mobile sensing process and their presence reduces the trustworthiness of the model output to end users and introduces uncertainty in the estimation process. Quantifying this uncertainty level is necessary for exploring the practical utility of these streams in representing the health and wellness states of individuals in natural environments. The smoking risk prediction model is another ML-based model dependent on these streams and their associated uncertainties. Future research works can output a measure of uncertainty alongside the lapse risk score. An uncertainty measure can be more helpful to decide what interventions to deliver and when to deliver them based on the risk score.

To increase the number of labeled instances of at-risk moments, future works can utilize other sources of labels. EMAs offer a possible source of such noisy labels. EMAs are almost ubiquitous in smoking cessation studies. Using EMAs participants self-report smoking lapses. They fill in structured questionnaires designed to locate the lapse timing within a block of the exact time of smoking. For example, the participants may report that they smoked within the last 2 hours. Thus, the lapse-reports obtained using EMAs come with a coarse and imprecise timing resolution. Future works can develop novel methodologies for utilizing these label sources. Future works can also employ novel study designs to collect ground truth labels for low-risk moments as well.

Our proposed risk estimation model achieves a reasonable performance(IHR) using only the stress, location, and activity features. Future works can boost the performance further by supplementing them with craving, self-efficacy, presence of other

cues such as noisy locations, graffiti, and other situational indicators that may affect the risk of lapse.

We evaluate the utility of the developed risk estimation models using simulation experiments of delivering smoking interventions as informed by the smoking lapse risk scores. A comprehensive evaluation of the developed methods will require design of a randomized clinical trial (RCT) to quantify the performance of the proposed models. Future works can employ adapt the developed methodologies for deployment in a RCT.

Finally, a novel future application can use the developed risk scores to dictate the delivery of Ecological Momentary Assessments based on the smoking lapse risk scores. In a smoking cessation research study, researchers employ EMAs to assess participants' health and wellness states of individuals at regular intervals. EMAs are typically delivered to participants in regular blocks throughout the day. Filling in EMA questionnaires requires time and effort, thus putting a burden on the study participants willing to quit smoking. Using the smoking lapse risk scores to inform the delivery of EMAs can improve the effectiveness of the delivered EMAs and reduce the number of EMAs delivered daily.

REFERENCES

[1] S. Akther, N. Saleheen, S. A. Samiei, V. Shetty, E. Ertin, and S. Kumar, "moral: An mhealth model for inferring oral hygiene behaviors in-the-wild using wrist-worn inertial sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 1, pp. 1–25, 2019.

[2] Jon, "Animation – Walking character blocking pass," 6 2013. [Online]. Available: https://jwong83.wordpress.com/2013/05/08/animation-walking-character-blocking-pass/

[3] "World health organization tobacco fact sheet." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/tobacco

[4] X. Dai, E. Gakidou, and A. D. Lopez, "Evolution of the global smoking epidemic over the past half century: strengthening the evidence base for policy action," *Tobacco Control*, vol. 31, no. 2, pp. 129–137, 2022.

[5] R. Peto and A. D. Lopez, "The future worldwide health effects of current smoking patterns," *Tobacco and public health: Science and policy*, vol. 281, no. 6, pp. 281–286, 2004.

[6] "Smoking cessation: Fast facts," Mar 2022. [Online]. Available: https://www.cdc.gov/tobacco/data_statistics/fact_sheets/cessation/smoking-cessation-fast-facts/index.html

[7] M.-J. Yang, S. K. Sutton, L. M. Hernandez, S. R. Jones, D. W. Wetter, S. Kumar, and C. Vinci, "A just-in-time adaptive intervention (jitai) for smoking cessation: Feasibility and acceptability findings," *Addictive Behaviors*, vol. 136, p. 107467, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306460322002337

[8] I. Nahum-Shani, S. N. Smith, B. J. Spring, L. M. Collins, K. Witkiewitz, A. Tewari, and S. A. Murphy, "Just-in-time adaptive interventions (jitais) in mobile health: Key components and design principles for ongoing health behavior support." [Online]. Available: https://academic.oup.com/abm/article/52/6/446/4733473

[9] S. Consolvo, D. W. McDonald, T. Toscos, M. Y. Chen, J. Froehlich, B. Harrison, P. Klasnja, A. LaMarca, L. LeGrand, R. Libby, *et al.*, "Activity sensing in the wild: a field trial of ubifit garden," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2008, pp. 1797–1806.

[10] N. Saleheen, A. A. Ali, S. M. Hossain, H. Sarker, S. Chatterjee, B. Marlin, E. Ertin, M. al'Absi, and S. Kumar, "puffmarker: a multi-sensor approach for pinpointing the timing of first lapse in smoking cessation," in *ACM UbiComp*, 2015, pp. 999–1010.

[11] J.-K. Min, A. Doryab, J. Wiese, S. Amini, J. Zimmerman, and J. I. Hong, "Toss'n'turn: smartphone as sleep and sleep quality detector," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2014, pp. 477–486.

[12] S. Bae, D. Ferreira, B. Suffoletto, J. C. Puyana, R. Kurtz, T. Chung, and A. K. Dey, "Detecting drinking episodes in young adults using smartphone-based sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 2, pp. 1–36, 2017.

[13] R. Bari, R. J. Adams, M. M. Rahman, M. B. Parsons, E. H. Buder, and S. Kumar, "rconverse: Moment by moment conversation detection using a mobile respiration sensor," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 2, no. 1, pp. 1–27, 2018.

[14] E. Thomaz, I. Essa, and G. D. Abowd, "A practical approach for recognizing eating moments with wrist-mounted inertial sensing," in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2015, pp. 1029–1040.

[15] K. Hovsepian, M. al'Absi, E. Ertin, T. Kamarck, M. Nakajima, and S. Kumar, "cstress: towards a gold standard for continuous stress assessment in the mobile environment," in *ACM UbiComp*, 2015, pp. 493–504.

[16] R. Wang, W. Wang, A. DaSilva, J. F. Huckins, W. M. Kelley, T. F. Heatherton, and A. T. Campbell, "Tracking depression dynamics in college students using mobile phone and wearable sensing," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–26, 2018.

[17] R. Wang, M. S. Aung, S. Abdullah, R. Brian, A. T. Campbell, T. Choudhury, M. Hauser, J. Kane, M. Merrill, E. A. Scherer, *et al.*, "Crosscheck: toward passive sensing and detection of mental health changes in people with schizophrenia," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 886–897.

[18] L. Canzian and M. Musolesi, "Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis," in *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 2015, pp. 1293–1304.

[19] M.-Z. Poh, T. Loddenkemper, N. C. Swenson, S. Goyal, J. R. Madsen, and R. W. Picard, "Continuous monitoring of electrodermal activity during epileptic seizures using a wearable sensor," in *IEEE EMBC*, 2010, pp. 4415–4418.

[20] S. Chatterjee, M. M. Rahman, T. Ahmed, N. Saleheen, E. Nemati, V. Nathan, K. Vatanparvar, and J. Kuang, "Assessing severity of pulmonary obstruction from respiration phase-based wheeze-sensing using mobile sensors," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.

[21] Y.-Y. Liu, S. Li, F. Li, L. Song, and J. M. Rehg, "Efficient learning of continuous-time hidden markov models for disease progression," in *Advances in neural information processing systems*, 2015, pp. 3600–3608.

[22] S. L. Kenford, M. C. Fiore, D. E. Jorenby, S. S. Smith, D. Wetter, and T. B. Baker, "Predicting smoking cessation: who will quit with and without the nicotine patch," *Jama*, vol. 271, no. 8, pp. 589–594, 1994.

[23] S. Shiffman, J. A. Paty, M. Gnys, J. A. Kassel, and M. Hickcox, "First lapses to smoking: within-subjects analysis of real-time reports." *Journal of consulting and clinical psychology*, vol. 64, no. 2, p. 366, 1996.

[24] S. Chatterjee, "Machine learning models for predicting the imminent risk of impulsive behaviors using mhealth sensors," 2021.

[25] M. E. Larimer and G. A. Marlatt, "Relapse prevention: An overview of marlatt's cognitive-behavioral model," *Psychosocial treatments*, pp. 11–28, 2004.

[26] J. S. Baer, T. Karmack, E. Lichtenstein, and C. C. Ransom, "Prediction of smoking relapse: analyses of temptations and transgressions after initial cessation." *Journal of consulting and clinical psychology*, vol. 57, no. 5, p. 623, 1989.

[27] J. S. Baer and E. Lichtenstein, "Classification and prediction of smoking relapse episodes: an exploration of individual differences." *Journal of consulting and clinical psychology*, vol. 56, no. 1, p. 104, 1988.

[28] K. A. O'Connell and E. J. Martin, "Highly tempting situations associated with abstinence, temporary lapse, and relapse among participants in smoking cessation programs." *Journal of consulting and clinical psychology*, vol. 55, no. 3, p. 367, 1987.

[29] S. Shiffman, "Relapse following smoking cessation: a situational analysis." *Journal of consulting and clinical psychology*, vol. 50, no. 1, p. 71, 1982.

[30] S. Chatterjee, K. Hovsepian, H. Sarker, N. Saleheen, M. al'Absi, G. Atluri, E. Ertin, C. Lam, A. Lemieux, M. Nakajima, *et al.*, "mcrave: Continuous estimation of craving during smoking cessation," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2016, pp. 863–874.

[31] R. Borland, "Slip-ups and relapse in attempts to quit smoking," *Addictive behaviors*, vol. 15, no. 3, pp. 235–245, 1990.

[32] S. Chatterjee, A. Moreno, S. L. Lizotte, S. Akther, E. Ertin, C. P. Fagundes, C. Lam, J. M. Rehg, N. Wan, D. W. Wetter, *et al.*, "Smokingopp: Detecting the smoking'opportunity'context using mobile sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 1, pp. 1–26, 2020.

[33] J. Yoon, A. Alaa, S. Hu, and M. Schaar, "Forecasticu: a prognostic decision support system for timely prediction of intensive care unit admission," in *International Conference on Machine Learning*. PMLR, 2016, pp. 1680–1689.

[34] Z. Zhao, A. Chen, W. Hou, J. M. Graham, H. Li, P. S. Richman, H. C. Thode, A. J. Singer, and T. Q. Duong, "Prediction model and risk scores of icu admission and mortality in covid-19," *PloS one*, vol. 15, no. 7, p. e0236618, 2020.

[35] S. Bhattacharya, V. Rajan, and H. Shrivastava, "Icu mortality prediction: a classification algorithm for imbalanced datasets," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[36] M. Ghassemi, M. Pimentel, T. Naumann, T. Brennan, D. Clifton, P. Szolovits, and M. Feng, "A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 29, no. 1, 2015.

[37] Y. Gao, G.-Y. Cai, W. Fang, H.-Y. Li, S.-Y. Wang, L. Chen, Y. Yu, D. Liu, S. Xu, P.-F. Cui, *et al.*, "Machine learning based early warning system enables accurate mortality risk prediction for covid-19," *Nature communications*, vol. 11, no. 1, pp. 1–10, 2020.

[38] H. Suresh, J. J. Gong, and J. V. Guttag, "Learning tasks for multitask learning: Heterogeneous patient populations in the icu," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 802–810.

[39] S. L. Hyland, M. Faltys, M. Hüser, X. Lyu, T. Gumbsch, C. Esteban, C. Bock, M. Horn, M. Moor, B. Rieck, *et al.*, "Early prediction of circulatory failure in the intensive care unit using machine learning," *Nature medicine*, vol. 26, no. 3, pp. 364–373, 2020.

[40] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," *Journal of the American Medical Informatics Association*, vol. 24, no. 2, pp. 361–370, 2017.

[41] T. Bai, S. Zhang, B. L. Egleston, and S. Vucetic, "Interpretable representation learning for healthcare via capturing disease progression through time," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 43–51.

[42] J. Futoma, S. Hariharan, K. Heller, M. Sendak, N. Brajer, M. Clement, A. Bedoya, and C. O'Brien, "An improved multi-output gaussian process rnn with real-time validation for early sepsis detection," in *Machine Learning for Healthcare Conference*. PMLR, 2017, pp. 243–254.

[43] Y. Xu, S. Biswal, S. R. Deshpande, K. O. Maher, and J. Sun, "Raim: Recurrent attentive and intensive model of multimodal patient monitoring data," in *Proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*, 2018, pp. 2565–2573.

[44] B. Singh Walia, Q. Hu, J. Chen, F. Chen, J. Lee, N. Kuo, P. Narang, J. Batts, G. Arnold, and M. Madaio, "A dynamic pipeline for spatio-temporal fire risk prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 764–773.

[45] Q. Wang, J. Zhang, B. Guo, Z. Hao, Y. Zhou, J. Sun, Z. Yu, and Y. Zheng, "Cityguard: citywide fire risk forecasting using a machine learning approach," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 4, pp. 1–21, 2019.

[46] M. Madaio, S.-T. Chen, O. L. Haimson, W. Zhang, X. Cheng, M. Hinds-Aldrich, D. H. Chau, and B. Dilkina, "Firebird: Predicting fire risk and prioritizing fire inspections in atlanta," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 185–194.

[47] S. Moosavi, M. H. Samavatian, S. Parthasarathy, R. Teodorescu, and R. Ramnath, "Accident risk prediction based on heterogeneous sparse data: New dataset and insights," in *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2019, pp. 33–42.

[48] Q. Chen, X. Song, H. Yamada, and R. Shibasaki, "Learning deep representation from big and heterogeneous data for traffic accident inference," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[49] S. Moosavi, M. H. Samavatian, A. Nandi, S. Parthasarathy, and R. Ramnath, "Short and long-term pattern discovery over large-scale geo-spatiotemporal data," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2905–2913.

[50] A. Mosavi, P. Ozturk, and K.-w. Chau, "Flood prediction using machine learning models: Literature review," *Water*, vol. 10, no. 11, p. 1536, 2018.

[51] M. Salehi, L. I. Rusu, T. Lynar, and A. Phan, "Dynamic and robust wildfire risk prediction system: an unsupervised approach," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 245–254.

[52] S. Gholami, N. Kodandapani, J. Wang, and J. L. Ferres, "Where there's smoke, there's fire: Wildfire risk predictive modeling via historical climate data," in *Annual Conference on Innovative Applications of Artificial Intelligence (IAAI)*, 2021.

[53] R. E. Bliss, A. J. Garvey, J. W. Heinold, and J. L. Hitchcock, "The influence of situation and coping on relapse crisis outcomes after smoking cessation." *Journal of consulting and clinical psychology*, vol. 57, no. 3, p. 443, 1989.

[54] A. A. Stone and S. Shiffman, "Ecological momentary assessment (ema) in behavorial medicine." *Annals of Behavioral Medicine*, 1994.

[55] A. A. Ali, S. M. Hossain, K. Hovsepian, M. M. Rahman, K. Plarre, and S. Kumar, "mpuff: automated detection of cigarette smoking puffs from respiration measurements," in *Proceedings of the 11th international conference on Information Processing in Sensor Networks*, 2012, pp. 269–280.

[56] R. I. Ramos-Garcia, S. Tiffany, and E. Sazonov, "Using respiratory signals for the recognition of human activities," in *2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2016, pp. 173–176.

[57] B. R. Raiff, Ç. Karataş, E. A. McClure, D. Pompili, and T. A. Walls, "Laboratory validation of inertial body sensors to detect cigarette smoking arm movements," *Electronics*, vol. 3, no. 1, pp. 87–110, 2014.

[58] A. Parate, M.-C. Chiu, C. Chadowitz, D. Ganesan, and E. Kalogerakis, "Risq: Recognizing smoking gestures with inertial sensors on a wristband," in *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. ACM, 2014, pp. 149–161.

[59] Q. Tang, *Automated detection of puffing and smoking with wrist accelerometers*. Northeastern University, 2014.

[60] M. Shoaib, H. Scholten, P. J. Havinga, and O. D. Incel, "A hierarchical lazy smoking detection algorithm using smartwatch sensors," in *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 2016, pp. 1–6.

[61] V. Senyurek, M. Imtiaz, P. Belsare, S. Tiffany, and E. Sazonov, "Cigarette smoking detection with an inertial sensor and a smart lighter," *Sensors*, vol. 19, no. 3, p. 570, 2019.

[62] V. Y. Senyurek, M. H. Imtiaz, P. Belsare, S. Tiffany, and E. Sazonov, "A cnn-lstm neural network for recognition of puffing in smoking episodes using wearable sensors," *Biomedical Engineering Letters*, vol. 10, pp. 195–203, 2020.

[63] F. Mordelet and J.-P. Vert, "A bagging svm to learn from positive and unlabeled examples," *Pattern Recognition Letters*, vol. 37, pp. 201–209, 2014.

[64] E. Sheetrit, N. Nissim, D. Klimov, and Y. Shahar, "Temporal probabilistic profiles for sepsis prediction in the icu," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2961–2969.

[65] J. Futoma, S. Hariharan, and K. Heller, "Learning to detect sepsis with a multitask gaussian process rnn classifier," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1174–1182.

[66] C. J. Gwaltney, S. Shiffman, M. H. Balabanis, and J. A. Paty, "Dynamic self-efficacy and outcome expectancies: prediction of smoking lapse and relapse." *Journal of abnormal psychology*, vol. 114, no. 4, p. 661, 2005.

[67] S. Shiffman and A. J. Waters, "Negative affect and smoking lapses: a prospective analysis." *Journal of consulting and clinical psychology*, vol. 72, no. 2, p. 192, 2004.

[68] S. Shiffman, M. H. Balabanis, C. J. Gwaltney, J. A. Paty, M. Gnys, J. D. Kassel, M. Hickcox, and S. M. Paton, "Prediction of lapse from associations between smoking and situational antecedents assessed by ecological momentary assessment," *Drug and alcohol dependence*, vol. 91, no. 2-3, pp. 159–168, 2007.

[69] T.-T. Phan, S. Muralidhar, and D. Gatica-Perez, "Drinks & crowds: Characterizing alcohol consumption through crowdsensing and social media," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 2, pp. 1–30, 2019.

[70] B. T. Gullapalli, A. Natarajan, G. A. Angarita, R. T. Malison, D. Ganesan, and T. Rahman, "On-body sensing of cocaine craving, euphoria and drug-seeking behavior using cardiac and respiratory signals," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 2, pp. 1–31, 2019.

[71] F. Gu, M.-H. Chung, M. Chignell, S. Valaee, B. Zhou, and X. Liu, "A survey on deep learning for human activity recognition," *ACM Computing Surveys (CSUR)*, vol. 54, no. 8, pp. 1–34, 2021.

[72] E. Ertin, N. Stohs, S. Kumar, A. Raij, M. al'Absi, and S. Shah, "Autosense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field," in *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*. ACM, 2011, pp. 274–287.

[73] W. Z. Tee, R. Dave, J. Seliya, and M. Vanamala, "A close look into human activity recognition models using deep learning," in *2022 3rd International Conference on Computing, Networks and Internet of Things (CNIOT)*. IEEE, 2022, pp. 201–206.

[74] Y. Tang, Q. Teng, L. Zhang, F. Min, and J. He, "Layer-wise training convolutional neural networks with smaller filters for human activity recognition using wearable sensors," *IEEE Sensors Journal*, vol. 21, no. 1, pp. 581–592, 2020.

[75] W. Qi, H. Su, C. Yang, G. Ferrigno, E. De Momi, and A. Aliverti, "A fast and robust deep convolutional neural networks for complex human activity recognition using smartphone," *Sensors*, vol. 19, no. 17, p. 3731, 2019.

[76] G. M. Weiss, "Wisdm smartphone and smartwatch activity and biometrics dataset," *UCI Machine Learning Repository: WISDM Smartphone and Smartwatch Activity and Biometrics Dataset Data Set*, vol. 7, pp. 133 190–133 202, 2019.

[77] A. Temko, "Estimation of heart rate from photoplethysmography during physical exercise using wiener filtering and the phase vocoder," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 1500–1503.

[78] Z. Zhang, "Photoplethysmography-based heart rate monitoring in physical activities via joint sparse spectrum reconstruction," *IEEE transactions on biomedical engineering*, vol. 62, no. 8, pp. 1902–1910, 2015.

[79] H. Lee, H. Chung, and J. Lee, "Motion artifact cancellation in wearable photoplethysmography using gyroscope," *IEEE Sensors Journal*, vol. 19, no. 3, pp. 1166–1175, 2018.

[80] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, "Deep ppg: Large-scale heart rate estimation with convolutional neural networks," *Sensors*, vol. 19, no. 14, p. 3079, 2019.

[81] A. Reiss, P. Schmidt, I. Indlekofer, and K. Van Laerhoven, "Ppg-based heart rate estimation with time-frequency spectra: A deep learning approach," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM, 2018, pp. 1283–1292.

[82] S. Salehizadeh, D. Dao, J. Bolkhovsky, C. Cho, Y. Mendelson, and K. H. Chon, "A novel time-varying spectral filtering algorithm for reconstruction of motion artifact corrupted heart rate signals during intense physical activities using a wearable photoplethysmogram sensor," *Sensors*, vol. 16, no. 1, p. 10, 2016.

[83] S. Salehizadeh, D. Dao, J. Bolkhovsky, C. Cho, Y. Mendelson, and K. Chon, "A novel time-varying spectral filtering algorithm for reconstruction of motion artifact corrupted heart rate signals during intense physical activities using a wearable photoplethysmogram sensor," *Sensors*, vol. 16, no. 1, p. 10, 2015.

[84] J. Park, S. Lee, and M. Jeon, "Atrial fibrillation detection by heart rate variability in poincare plot," *Biomedical engineering online*, vol. 8, no. 1, p. 38, 2009.

[85] M. Zabihi, A. B. Rad, A. K. Katsaggelos, S. Kiranyaz, S. Narkilahti, and M. Gabbouj, "Detection of atrial fibrillation in ecg hand-held devices using a random forest classifier," in *2017 Computing in Cardiology (CinC)*. IEEE, 2017, pp. 1–4.

[86] S. M. Hossain, A. A. Ali, M. M. Rahman, E. Ertin, D. Epstein, A. Kennedy, K. Preston, A. Umbricht, Y. Chen, and S. Kumar, "Identifying drug (cocaine) intake events from acute physiological response in the presence of free-living physical activity," in *Proceedings of the 13th international symposium on Information processing in sensor networks*. IEEE Press, 2014, pp. 71–82.

[87] A. Natarajan, G. Angarita, E. Gaiser, R. Malison, D. Ganesan, and B. M. Marlin, "Domain adaptation methods for improving lab-to-field generalization of cocaine detection using wearable ecg," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 875–885.

[88] A. Natarajan, A. Parate, E. Gaiser, G. Angarita, R. Malison, B. Marlin, and D. Ganesan, "Detecting cocaine use with wearable electrocardiogram sensors," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, 2013, pp. 123–132.

[89] A. H. Khandoker, M. Palaniswami, and C. K. Karmakar, "Support vector machines for automated recognition of obstructive sleep apnea syndrome from ecg recordings," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 1, pp. 37–48, Jan 2009.

[90] P. Melillo, C. Formisano, U. Bracale, and L. Pecchia, "Classification tree for real-life stress detection using linear heart rate variability analysis. case study: students under stress due to university examination," in *World Congress on Medical Physics and Biomedical Engineering May 26-31, 2012, Beijing, China*. Springer, 2013, pp. 477–480.

[91] G. Mark, Y. Wang, and M. Niiya, "Stress and multitasking in everyday college life: an empirical study of online activity," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2014, pp. 41–50.

[92] F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, and M. Griss, "Activity-aware mental stress detection using physiological sensors," in *International conference on Mobile computing, applications, and services*. Springer, 2010, pp. 282–301.

[93] V. Mishra, G. Pope, S. Lord, S. Lewia, B. Lowens, K. Caine, S. Sen, R. Halter, and D. Kotz, "The case for a commodity hardware solution for stress detection," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. ACM, 2018, pp. 1717–1728.

[94] ——, "Continuous detection of physiological stress with commodity hardware," *ACM Transactions on Computing for Healthcare*, vol. 1, no. 2, pp. 1–30, 2020.

[95] X. Sun, P. Yang, and Y.-T. Zhang, "Assessment of photoplethysmogram signal quality using morphology integrated with temporal information approach," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2012, pp. 3456–3459.

[96] N. Sharma and T. Gedeon, "Computational models of stress in reading using physiological and physical sensor data," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2013, pp. 111–122.

[97] J. Hong, J. Ramos, and A. Dey, "Understanding physiological responses to stressors during physical activity," in *ACM UbiComp*, 2012, pp. 270–279.

[98] J. Choi, B. Ahmed, and R. Gutierrez-Osuna, "Development and evaluation of an ambulatory stress monitor based on wearable sensors," *IEEE transactions on information technology in biomedicine*, vol. 16, no. 2, pp. 279–286, 2012.

[99] J. Zhai and A. Barreto, "Stress detection in computer users based on digital signal processing of noninvasive physiological variables," in *2006 international conference of the IEEE engineering in medicine and biology society*. IEEE, 2006, pp. 1355–1358.

[100] T. Rahman, M. Zhang, S. Voida, and T. Choudhury, "Towards accurate non-intrusive recollection of stress levels using mobile sensing and contextual recall," in *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*, 2014, pp. 166–169.

[101] N. Pinheiro, R. Couceiro, J. Henriques, J. Muehlsteff, I. Quintal, L. Goncalves, and P. Carvalho, "Can ppg be used for hrv analysis?" in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 2945–2949.

[102] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 2018, pp. 400–408.

[103] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "Continuous stress detection using a wrist device: in laboratory and real life," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 2016, pp. 1185–1193.

[104] Y. Cho, S. J. Julier, and N. Bianchi-Berthouze, "Instant stress: Detection of perceived mental stress through smartphone photoplethysmography and thermal imaging," *JMIR mental health*, vol. 6, no. 4, p. e10140, 2019.

[105] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," *Journal of biomedical informatics*, vol. 73, pp. 159–170, 2017.

[106] Y. S. Can, N. Chalabianloo, D. Ekiz, and C. Ersoy, "Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study," *Sensors*, vol. 19, no. 8, p. 1849, 2019.

[107] H.-W. Lee, J.-W. Lee, W.-G. Jung, and G.-K. Lee, "The periodic moving average filter for removing motion artifacts from ppg signals," *International Journal of Control, Automation, and Systems*, vol. 5, no. 6, pp. 701–706, 2007.

[108] H. Han and J. Kim, "Artifacts in wearable photoplethysmographs during daily life motions and their reduction with least mean square based active noise cancellation method," *Computers in biology and medicine*, vol. 42, no. 4, pp. 387–393, 2012.

[109] P. Wei, R. Guo, J. Zhang, and Y. Zhang, "A new wristband wearable sensor using adaptive reduction filter to reduce motion artifact," in *2008 International Conference on Information Technology and Applications in Biomedicine*. IEEE, 2008, pp. 278–281.

[110] P. K. Lim, S.-C. Ng, N. H. Lovell, Y. P. Yu, M. P. Tan, D. McCombie, E. Lim, and S. J. Redmond, "Adaptive template matching of photoplethysmogram pulses to detect motion artefact," *Physiological measurement*, vol. 39, no. 10, p. 105005, 2018.

[111] M. Raghuram, K. V. Madhav, E. H. Krishna, N. R. Komalla, K. Sivani, and K. A. Reddy, "Dual-tree complex wavelet transform for motion artifact reduction of ppg signals," in *2012 IEEE international symposium on medical measurements and applications proceedings*. IEEE, 2012, pp. 1–4.

[112] C. Lee and Y. T. Zhang, "Reduction of motion artifacts from photoplethysmographic recordings using a wavelet denoising approach," in *IEEE EMBS Asian-Pacific Conference on Biomedical Engineering, 2003.* IEEE, 2003, pp. 194–195.

[113] M. Raghuram, K. V. Madhav, E. H. Krishna, N. R. Komalla, K. Sivani, and K. A. Reddy, "Hht based signal decomposition for reduction of motion artifacts in photoplethysmographic signals," in *2012 IEEE International Instrumentation and Measurement Technology Conference Proceedings*. IEEE, 2012, pp. 1730–1734.

[114] Q. Wang, P. Yang, and Y. Zhang, "Artifact reduction based on empirical mode decomposition (emd) in photoplethysmography for pulse rate detection," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. IEEE, 2010, pp. 959–962.

[115] Y. Patil, S. Tiffany, and E. Sazonov, "Understanding smoking behavior using wearable sensors: Relative importance of various sensor modalities," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Ieee, 2014, pp. 6899–6902.

[116] R. Alharbi, S. Shahi, S. Cruz, L. Li, S. Sen, M. Pedram, C. Romano, J. Hester, A. K. Katsaggelos, and N. Alshurafa, "Smokemon: Unobtrusive extraction of smoking topography using wearable energy-efficient thermal," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 4, pp. 1–25, 2023.

[117] A. Yurtman and B. Barshan, "Activity recognition invariant to sensor orientation with wearable motion sensors," *Sensors*, vol. 17, no. 8, p. 1838, 2017.

[118] M. Straczkiewicz, N. W. Glynn, and J. Harezlak, "On placement, location and orientation of wrist-worn tri-axial accelerometers during free-living measurements," *Sensors*, vol. 19, no. 9, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/9/2095

[119] A. Yurtman, B. Barshan, and B. Fidan, "Activity recognition invariant to wearable sensor unit orientation using differential rotational transformations represented by quaternions," *Sensors*, vol. 18, no. 8, 2018. [Online]. Available: https://www.mdpi.com/1424-8220/18/8/2725

[120] O. Banos, M. A. Toth, M. Damas, H. Pomares, and I. Rojas, "Dealing with the effects of sensor displacement in wearable activity recognition," *Sensors*, vol. 14, no. 6, pp. 9995–10 023, 2014.

[121] S. Thiemjarus, "A device-orientation independent method for activity recognition," in *2010 International Conference on Body Sensor Networks*. IEEE, 2010, pp. 19–23.

[122] Y. E. Ustev, O. Durmaz Incel, and C. Ersoy, "User, device and orientation independent human activity recognition on mobile phones: Challenges and a proposal," in *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, ser. UbiComp '13 Adjunct. New York, NY, USA: Association for Computing Machinery, 2013, p. 1427–1436. [Online]. Available: https://doi.org/10.1145/2494091.2496039

[123] K. Kunze and P. Lukowicz, "Sensor placement variations in wearable activity recognition," *IEEE Pervasive Computing*, vol. 13, no. 4, pp. 32–41, 2014.

[124] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava, "Using mobile phones to determine transportation modes," *ACM Transactions on Sensor Networks (TOSN)*, vol. 6, no. 2, pp. 1–27, 2010.

[125] S. Bhattacharya, P. Nurmi, N. Hammerla, and T. Plötz, "Using unlabeled data in a sparse-coding framework for human activity recognition," *Pervasive and Mobile Computing*, vol. 15, pp. 242–262, 2014.

[126] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. Havinga, "Fusion of smartphone motion sensors for physical activity recognition," *Sensors*, vol. 14, no. 6, pp. 10 146–10 176, 2014.

[127] D. De, P. Bharti, S. K. Das, and S. Chellappan, "Multimodal wearable sensing for fine-grained activity recognition in healthcare," *IEEE Internet Computing*, vol. 19, no. 5, pp. 26–35, 2015.

[128] T. Hur, J. Bang, D. Kim, O. Banos, and S. Lee, "Smartphone location-independent physical activity recognition based on transportation natural vibration analysis," *Sensors*, vol. 17, no. 4, p. 931, 2017.

[129] H. Lu, J. Yang, Z. Liu, N. D. Lane, T. Choudhury, and A. T. Campbell, "The jigsaw continuous sensing engine for mobile phone applications," in *Proceedings of the 8th ACM conference on embedded networked sensor systems*, 2010, pp. 71–84.

[130] J. Yang, "Toward physical activity diary: motion recognition using simple acceleration features with mobile phones," in *Proceedings of the 1st international workshop on Interactive multimedia for consumer electronics*, 2009, pp. 1–10.

[131] J. Morales, D. Akopian, and S. Agaian, "Human activity recognition by smartphones regardless of device orientation," in *Mobile Devices and Multimedia: Enabling Technologies, Algorithms, and Applications 2014*, vol. 9030. SPIE, 2014, pp. 134–145.

[132] J. Bekker and J. Davis, "Learning from positive and unlabeled data: A survey," *Machine Learning*, vol. 109, no. 4, pp. 719–760, 2020.

[133] K. Jaskie and A. Spanias, "Positive and unlabeled learning algorithms and applications: A survey," in *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE, 2019, pp. 1–8.

[134] C. Elkan and K. Noto, "Learning classifiers from only positive and unlabeled data," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 213–220.

[135] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, "Building text classifiers using positive and unlabeled examples," in *Third IEEE international conference on data mining*. IEEE, 2003, pp. 179–186.

[136] W. S. Lee and B. Liu, "Learning with positive and unlabeled examples using weighted logistic regression."

[137] M. Claesen, F. D. Smet, J. A. K. Suykens, and B. D. Moor, "A robust ensemble approach to learn from positive and unlabeled data using svm base models," *Neurocomputing*, vol. 160, pp. 73–84, 2 2014. [Online]. Available: http://arxiv.org/abs/1402.3144http://dx.doi.org/10.1016/j.neucom.2014.10.081

[138] A. Acharya, S. Sanghavi, L. Jing, B. Bhushanam, D. Choudhary, M. Rabbat, and I. Dhillon, "Positive unlabeled contrastive learning," 6 2022. [Online]. Available: https://arxiv.org/abs/2206.01206v1

[139] H. G. Ramaswamy, C. Scott, and A. Tewari, "Mixture proportion estimation via kernel embedding of distributions," *33rd International Conference on Machine Learning, ICML 2016*, vol. 5, pp. 2996–3004, 3 2016. [Online]. Available: https://arxiv.org/abs/1603.02501v2

[140] D. Ivanov, "Dedpul: Difference-of-estimated-densities-based positive-unlabeled learning," 2019. [Online]. Available: https://arxiv.org/abs/1902.06965

[141] S. Garg, Y. Wu, A. Smola, S. Balakrishnan, and Z. C. Lipton, "Mixture proportion estimation and pu learning: A modern approach," 2021. [Online]. Available: https://arxiv.org/abs/2111.00980

[142] G. Niu and M. Sugiyama, "Convex formulation for learning from positive and unlabeled data marthinus christoffel du plessis."

[143] R. Kiryo, G. Niu, M. C. d. Plessis, and M. Sugiyama, "Positive-unlabeled learning with non-negative risk estimator," 2017. [Online]. Available: https://arxiv.org/abs/1703.00593

[144] B. Kayhan Tetik, I. Gedik Tekinemre, and S. Taş, "The effect of the covid-19 pandemic on smoking cessation success," *Journal of Community Health*, vol. 46, no. 3, pp. 471–475, 2021.

[145] F. Naughton, S. Hopewell, N. Lathia, R. Schalbroeck, C. Brown, C. Mascolo, A. McEwen, and S. Sutton, "A context-sensing mobile phone app (q sense) for smoking cessation: a mixed-methods study," *JMIR mHealth and uHealth*, vol. 4, no. 3, p. e106, 2016.

[146] I. Nahum-Shani, S. N. Smith, B. J. Spring, L. M. Collins, K. Witkiewitz, A. Tewari, and S. A. Murphy, "Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support," *Annals of Behavioral Medicine*, vol. 52, no. 6, pp. 446–462, 2018.

[147] H. Sarker, K. Hovsepian, S. Chatterjee, I. Nahum-Shani, S. A. Murphy, B. Spring, E. Ertin, M. Al'Absi, M. Nakajima, and S. Kumar, "From markers to interventions: The case of just-in-time stress intervention," in *Mobile health*. Springer, 2017, pp. 411–433.

[148] P. Klasnja, E. B. Hekler, S. Shiffman, A. Boruvka, D. Almirall, A. Tewari, and S. A. Murphy, "Microrandomized trials: An experimental design for developing just-in-time adaptive interventions." *Health Psychology*, vol. 34, no. S, p. 1220, 2015.

[149] M. M. Rahman, R. Bari, A. A. Ali, M. Sharmin, A. Raij, K. Hovsepian, S. M. Hossain, E. Ertin, A. Kennedy, D. H. Epstein, *et al.*, "Are we there yet?: Feasibility of continuous stress assessment via wireless physiological sensors," in *ACM BCB*, 2014, pp. 479–488.

[150] B. Liu, W. S. Lee, P. S. Yu, and X. Li, "Partially supervised classification of text documents," in *ICML*, vol. 2, no. 485. Sydney, NSW, 2002, pp. 387–394.

[151] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st international conference on neural information processing systems*, 2017, pp. 4768–4777.

[152] *CDC: Smoking is the leading cause of preventable death*, Accessed April, 2019. [Online]. Available: https://www.cdc.gov/tobacco/data_statistics/fact_sheets/fast_facts/index.htm

[153] L. R. Reitzel, E. K. Cromley, Y. Li, Y. Cao, R. Dela Mater, C. A. Mazas, L. Cofta-Woerpel, P. M. Cinciripini, and D. W. Wetter, "The effect of tobacco outlet density and proximity on smoking cessation," *American Journal of Public Health*, vol. 101, no. 2, pp. 315–320, 2011.

[154] S. Shiffman, "Reflections on smoking relapse research," *Drug and alcohol review*, vol. 25, no. 1, pp. 15–20, 2006.

[155] S. Shiffman, M. Hickcox, J. A. Paty, M. Gnys, J. D. Kassel, and T. J. Richards, "Progression from a smoking lapse to relapse: prediction from abstinence violation effects, nicotine dependence, and lapse characteristics." *Journal of consulting and clinical psychology*, vol. 64, no. 5, p. 993, 1996.

[156] S. M. Hossain, T. Hnat, N. Saleheen, N. J. Nasrin, J. Noor, B.-J. Ho, T. Condie, M. Srivastava, and S. Kumar, "mcerebrum: A mobile sensing software platform for development and validation of digital biomarkers and interventions," in *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*. ACM, 2017, p. 7.

[157] T. Hnat, S. M. Hossain, N. Ali, S. Carini, T. Condie, I. Sim, M. B. Srivastava, and S. Kumar, "mcerebrum and cerebral cortex: A real-time collection, analytic, and intervention platform for high-frequency mobile sensor data." in *AMIA*, 2017.

[158] T. R. Kirchner, S. Shiffman, and E. P. Wileyto, "Relapse dynamics during smoking cessation: recurrent abstinence violation effects and lapse-relapse progression." *Journal of abnormal psychology*, vol. 121, no. 1, p. 187, 2012.

[159] H. Suresh, N. Hunt, A. Johnson, L. A. Celi, P. Szolovits, and M. Ghassemi, "Clinical intervention prediction and understanding using deep networks," *arXiv preprint arXiv:1705.08498*, 2017.

[160] H. Sarker, M. Tyburski, M. Rahman, K. Hovsepian, M. Sharmin, D. H. Epstein, K. L. Preston, C. D. Furr-Holden, A. Milam, I. Nahum-Shani, M. al'Absi, and S. Kumar, "Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data," in *ACM CHI*, 2016.

[161] Y. Zheng, "Trajectory data mining: an overview," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 6, no. 3, p. 29, 2015.

[162] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, and J. Zhou, "Patient subtyping via time-aware lstm networks," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 65–74.

[163] Y. Suhara, Y. Xu, and A. Pentland, "Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 715–724.

[164] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, vol. 8, no. 1, pp. 1–12, 2018.

[165] J. Xu, J. Wang, M. Long, *et al.*, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[166] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of AAAI*, 2021.

[167] S. Hossain, A. Ali, M. Rahman, E. Ertin, D. Epstein, A. Kennedy, K. Preston, A. Umbricht, Y. Chen, and S. Kumar, "Identifying drug (cocaine) intake events from acute physiological response in the presence of free-living physical activity," in *ACM IPSN*, 2014, pp. 71–82.

[168] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[169] N. Saleheen, M. A. Ullah, S. Chakraborty, D. S. Ones, M. Srivastava, and S. Kumar, "Wristprint: Characterizing user re-identification risks from wrist-worn accelerometry data," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 2021, pp. 2807–2823.

[170] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

[171] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[172] A. Kapoor and E. Horvitz, "Experience sampling for building predictive user models: a comparative study," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2008, pp. 657–666.

[173] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[174] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," 2019.

[175] H. Chen, S. Lundberg, and S.-I. Lee, "Explaining models by propagating shapley values of local components," in *Explainable AI in Healthcare and Medicine*. Springer, 2021, pp. 261–270.

[176] F. Künzler, V. Mishra, J.-N. Kramer, D. Kotz, E. Fleisch, and T. Kowatsch, "Exploring the state-of-receptivity for mhealth interventions," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 4, pp. 1–27, 2019.

[177] H. Sarker, M. Sharmin, A. Ali, M. Rahman, R. Bari, S. Hossain, and S. Kumar, "Assessing the availability of users to engage in just-in-time intervention in the natural environment," in *ACM UbiComp*, 2014, pp. 909–920.

[178] D. Teichmann, J. Klopp, A. Hallmann, K. Schuett, S. Wolfart, and M. Teichmann, "Detection of acute periodontal pain from physiological signals," *Physiological measurement*, vol. 39, no. 9, p. 095007, 2018.

[179] Y. L. Yang, H. S. Seok, G.-J. Noh, B.-M. Choi, and H. Shin, "Postoperative pain assessment indices based on photoplethysmography waveform analysis," *Frontiers in Physiology*, vol. 9, p. 1199, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fphys.2018.01199

[180] J. A. Johnstone, P. A. Ford, G. Hughes, T. Watson, and A. T. Garrett, "Bioharness™ multivariable monitoring device: part. i: validity," *Journal of sports science & medicine*, vol. 11, no. 3, p. 400, 2012.

[181] N. Isakadze and S. S. Martin, "How useful is the smartwatch ecg?" *Trends in Cardiovascular Medicine*, 2019.

[182] S. Tedesco, M. Sica, A. Ancillao, S. Timmons, J. Barton, and B. O'Flynn, "Validity evaluation of the fitbit charge2 and the garmin vivosmart hr+ in free-living environments in an older adult cohort," *JMIR mHealth and uHealth*, vol. 7, no. 6, p. e13084, 2019.

[183] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological measurement*, vol. 28, no. 3, p. R1, 2007.

[184] J. A. Sukor, S. Redmond, and N. Lovell, "Signal quality measures for pulse oximetry through waveform morphology analysis," *Physiological measurement*, vol. 32, no. 3, p. 369, 2011.

[185] C. Fischer, B. Dömer, T. Wibmer, and T. Penzel, "An algorithm for real-time pulse waveform segmentation and artifact detection in photoplethysmograms," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 2, pp. 372–381, 2017.

[186] W. Karlen, K. I. shi Kobayashi, J. M. Ansermino, and G. A. Dumont, "Photoplethysmogram signal quality estimation using repeated gaussian filters and cross-correlation." *Physiological measurement*, vol. 33 10, pp. 1617–29, 2012.

[187] M. Pflugradt, B. Moeller, and R. Orglmeister, "Opra: A fast on-line signal quality estimator for pulsatile signals," *IFAC-PapersOnLine*, vol. 48, no. 20, pp. 459–464, 2015.

[188] Q. Li and G. D. Clifford, "Dynamic time warping and machine learning for signal quality assessment of pulsatile signals," *Physiological measurement*, vol. 33, no. 9, p. 1491, 2012.

[189] A.-G. Pielmuș, D. Osterland, T. Tigges, M. Klum, R. Orglmeister, A. Feldheiser, and O. Hunsicker, "Dynamic time warping of pulse wave curves," *Current Directions in Biomedical Engineering*, vol. 4, no. 1, pp. 371 – 374, 2018. [Online]. Available: https://www.degruyter.com/view/journals/cdbme/4/1/article-p371.xml

[190] M. Elgendi, "Optimal signal quality index for photoplethysmogram signals," *Bioengineering*, vol. 3, no. 4, p. 21, 2016.

[191] M. Hartmut Gehring, H. M. ME, and P. Schmucker, "The effects of motion artifact and low perfusion on the performance of a new generation of pulse oximeters in volunteers undergoing hypoxemia," *Respiratory Care*, vol. 47, no. 1, pp. 48–60, 2002.

[192] N. Selvaraj, Y. Mendelson, K. H. Shelley, D. G. Silverman, and K. H. Chon, "Statistical approach for the detection of motion/noise artifacts in photoplethysmogram," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2011, pp. 4972–4975.

[193] R. Krishnan, B. Natarajan, and S. Warren, "Two-stage approach for detection and reduction of motion artifacts in photoplethysmographic data," *IEEE transactions on biomedical engineering*, vol. 57, no. 8, pp. 1867–1876, 2010.

[194] Y. Liang, M. Elgendi, Z. Chen, and R. Ward, "An optimal filter for short photoplethysmogram signals," *Scientific data*, vol. 5, p. 180076, 2018.

[195] A. Aliamiri and Y. Shen, "Deep learning based atrial fibrillation detection using wearable photoplethysmography sensor," in *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, 2018, pp. 442–445.

[196] S. Nemati, M. M. Ghassemi, V. Ambai, N. Isakadze, O. Levantsevych, A. Shah, and G. D. Clifford, "Monitoring and detecting atrial fibrillation using wearable technology," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, pp. 3394–3397.

[197] T. Zhao, Y. Wang, J. Liu, Y. Chen, J. Cheng, and J. Yu, "Trueheart: Continuous authentication on wrist-worn wearables using ppg-based biometrics," 2020.

[198] D. J. Choi, M. S. Choi, and S. J. Kang, "A wearable device platform for the estimation of sleep quality using simultaneously motion tracking and pulse oximetry," in *2016 IEEE International Conference on Consumer Electronics (ICCE)*, 2016, pp. 49–50.

[199] T. Pereira, K. Gadhoumi, M. Ma, L. Xiuyun, R. Xiao, R. A. Colorado, K. J. Keenan, K. Meisel, and X. Hu, "A supervised approach to robust photoplethysmography quality assessment," *IEEE Journal of Biomedical and Health Informatics*, 2019.

[200] E. K. Naeinia, I. Azimib, A. M. Rahmania, P. Liljebergb, and N. Dutta, "A real-time ppg quality assessment approach for healthcare internet-of-things," *Procedia Computer Science*, vol. 151, pp. 551–558, 2019.

[201] N. Pradhan, S. Rajan, and A. Adler, "Evaluation of the signal quality of wrist-based photoplethysmography," *Physiological measurement*, 2019.

[202] E. Sabeti, N. Reamaroon, M. Mathis, J. Gryak, M. Sjoding, and K. Najarian, "Signal quality measure for pulsatile physiological signals using morphological features: Applications in reliability measure for pulse oximetry," *Informatics in Medicine Unlocked*, vol. 16, p. 100222, 2019.

[203] R. Banerjee, R. Vempada, K. M. Mandana, A. D. Choudhury, and A. Pal, "Identifying coronary artery disease from photoplethysmogram," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, ser. UbiComp '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1084–1088. [Online]. Available: https://doi.org/10.1145/2968219.2972712

[204] R. Zangróniz, A. Martínez-Rodrigo, M. López, J. Pastor, and A. Fernández-Caballero, "Estimation of mental distress from photoplethysmography," *Applied Sciences*, vol. 8, no. 1, p. 69, Jan 2018. [Online]. Available: http://dx.doi.org/10.3390/app8010069

[205] S. Fallet, M. Lemay, P. Renevey, C. Leupi, E. Pruvot, and J.-M. Vesin, "Can one detect atrial fibrillation using a wrist-type photoplethysmographic device?" *Medical & Biological Engineering & Computing*, vol. 57, pp. 1–11, 09 2018.

[206] A. Sološenko, A. Petrėnas, B. Paliakaitė, L. Sörnmo, and V. Marozas, "Detection of atrial fibrillation using a wrist-worn device," *Physiological Measurement*, vol. 40, no. 2, p. 025003, feb 2019. [Online]. Available: https://doi.org/10.1088%2F1361-6579%2Fab029c

[207] P. Liao, W. Dempsey, H. Sarker, S. M. Hossain, M. Al'Absi, P. Klasnja, and S. Murphy, "Just-in-time but not too much: Determining treatment timing in mobile health," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 2, no. 4, pp. 1–21, 2018.

[208] M. Al'Absi, S. Bongard, T. Buchanan, G. A. Pincomb, J. Licinio, and W. R. Lovallo, "Cardiovascular and neuroendocrine adjustment to public speaking and mental arithmetic stressors," *Psychophysiology*, vol. 34, no. 3, pp. 266–275, 1997.

[209] P. Welch, "The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms," *IEEE Transactions on audio and electroacoustics*, vol. 15, no. 2, pp. 70–73, 1967.

[210] J. Song, D. Li, X. Ma, G. Teng, and J. Wei, "A robust dynamic heart-rate detection algorithm framework during intense physical activities using photoplethysmographic signals," *Sensors*, vol. 17, no. 11, p. 2450, 2017.

[211] H. Gehring, C. Hornberger, H. Matz, E. Konecny, and P. Schmucker, "Original contributions-the effects of motion artifact and low perfusion on the performance of a new generation of pulse oximeters in volunteers undergoing hypoxemia," *Respiratory Care*, vol. 47, no. 1, pp. 48–60, 2002.

[212] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[213] T. Schäck, M. Muma, and A. M. Zoubir, "Computationally efficient heart rate estimation during physical exercise using photoplethysmographic signals," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 2478–2481.

[214] Z. Zhang, Z. Pi, and B. Liu, "Troika: A general framework for heart rate monitoring using wrist-type photoplethysmographic signals during intensive physical exercise," *IEEE Transactions on biomedical engineering*, vol. 62, no. 2, pp. 522–531, 2014.

[215] D. E. Hinkle, W. Wiersma, and S. G. Jurs, *Applied statistics for the behavioral sciences*. Houghton Mifflin College Division, 2003, vol. 663.

[216] K. Plarre, A. Raij, S. Guha, M. al'Absi, E. Ertin, and S. Kumar, "Automated detection of sensor detachments for physiological sensing in the wild," in *Wireless Health 2010*. ACM, 2010, pp. 216–217.

[217] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE Trans. Biomed. Eng*, vol. 32, no. 3, pp. 230–236, 1985.

[218] G. G. Berntson, K. S. Quigley, J. F. Jang, and S. T. Boysen, "An approach to artifact identification: Application to heart period data," *Psychophysiology*, vol. 27, no. 5, pp. 586–598, 1990.

[219] K. Plarre, A. Raij, S. Hossain, A. Ali, M. Nakajima, M. Al'absi, E. Ertin, T. Kamarck, S. Kumar, M. Scott, *et al.*, "Continuous inference of psychological stress from sensory measurements collected in the natural environment," in *IEEE/ACM IPSN*, 2011, pp. 97–108.

[220] N. R. Lomb, "Least-squares frequency analysis of unequally spaced data," *Astrophysics and space science*, vol. 39, no. 2, pp. 447–462, 1976.

[221] J. D. Scargle, "Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data," *The Astrophysical Journal*, vol. 263, pp. 835–853, 1982.

[222] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[223] A. V. Dorogush, V. Ershov, and A. Gulin, "Catboost: gradient boosting with categorical features support," *arXiv preprint arXiv:1810.11363*, 2018.

[224] J. T. Hancock and T. M. Khoshgoftaar, "Catboost for big data: an interdisciplinary review," *Journal of big data*, vol. 7, no. 1, pp. 1–45, 2020.

[225] Ł. Kobyliński and A. Przepiórkowski, "Definition extraction with balanced random forests," in *Advances in Natural Language Processing: 6th International Conference, GoTAL 2008 Gothenburg, Sweden, August 25-27, 2008 Proceedings*. Springer, 2008, pp. 237–247.

[226] C. for Disease Control, P. (CDC, *et al.*, "Smoking-attributable mortality, years of potential life lost, and productivity losses–united states, 2000-2004," *MMWR. Morbidity and mortality weekly report*, vol. 57, no. 45, pp. 1226–1228, 2008.

[227] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[228] Z. Fu, Z. Tian, Y. Xu, and C. Qiao, "A two-step clustering approach to extract locations from individual gps trajectory data," *ISPRS International Journal of Geo-Information*, vol. 5, no. 10, p. 166, 2016.

[229] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.