University of Missouri

PHYSIOLOGICAL DATA ANALYSIS

ALCOHOL DRINKING PREDICTION

USING STATISTICAL AND DEEP LEARNING METHODS

Master's Thesis Defense

Can Li

Advisor: Dr. Yi Shang

Contents

- Introduction
- Related Work
- Experiment Data
- Data Analysis Methods
- Experiment Results and Comparison
- Conclusion and Future Work

Contents

Introduction

Alcohol craving study based on real physiological data

1. Data was collected from mobile ambulatory assessment system
2. The type of sensor used is basis watch
3. The goal of this study is to predict whether people had drinking or not using machine learning pipeline

Problem Definition

Input: One dimensional skin temperature, heart rate, GSR(galvanic skin response) signal

Method: Data analysis pipeline

1. Data labeling
2. Data cleaning
3. Feature extraction
4. Classification

Output: {0, 1}, 0 is non-drinking and 1 is drinking

Motivation and Contributions

Motivation:

1. Previous work was doing drinking prediction based on each record. There is overlapping information in the result. Prediction based on drinking episode is more reasonable.
2. To try deep learning on drinking episode prediction

Contributions:

1. Came up with drinking episode and deep learning pipeline
2. New features were extracted
3. Found that heart rate is the most significant feature in drinking prediction
4. Achieve 88.89% accuracy for drinking episode prediction

Contents

Related Work

Hossain, Syed Monowar, et al. "Identifying drug (cocaine) intake events from acute physiological response in the presence of free-living physical activity." Proceedings of the 13th international symposium on Information processing in sensor networks. IEEE Press, 2014.

- This paper was identifying recovery time from cocaine intake, which gave me the idea to do drinking episode prediction

Related Work (cont'd)

Wergeles, Nickolas M. "AMD: Analysis of Mood Dysregulation A Machine Learning Approach" 2016.

1. He is doing mood dysregulation prediction from physiological data. My research is about drinking prediction.
2. Prediction is based on each 5-second record. My prediction is based on both 1-minute record and 30-minute data block.
3. Data cleaning method was introduced in his paper. I used the similar data cleaning method.

Related Work (cont'd)

Zhang, Chen. "Wearable Sensing Analysis – Identifying alcohol Drinking From Daily Physiological Data" 2016.

1. Doing alcohol drinking prediction on physiological data from SEM, Hexoskin sensors. My data is from basis watch.
2. His sample rate is 5 seconds. Mine is 1 minute.
3. Statistical features were extracted from 1-minute window. I extracted different statistical features and deep learning features based on 30-minute data block.

Contents

1. Data Overview

- Number of Users: 29
- Survey Data
  1) Initial Drinking
  2) Drinking Follow-ups
- Raw data (Sensor Data)
  - Sample rate: 1 minute
  - Features
    1) Skin Temperature
    2) Heart Rate
    3) GSR (galvanic skin response)

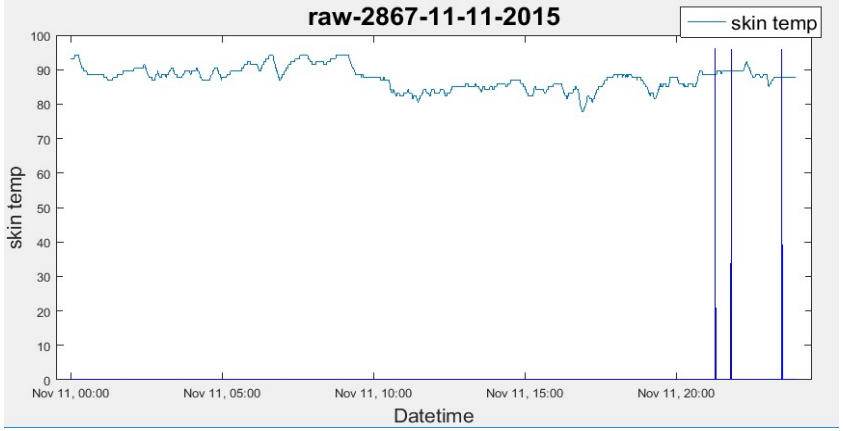
Survey Data Example

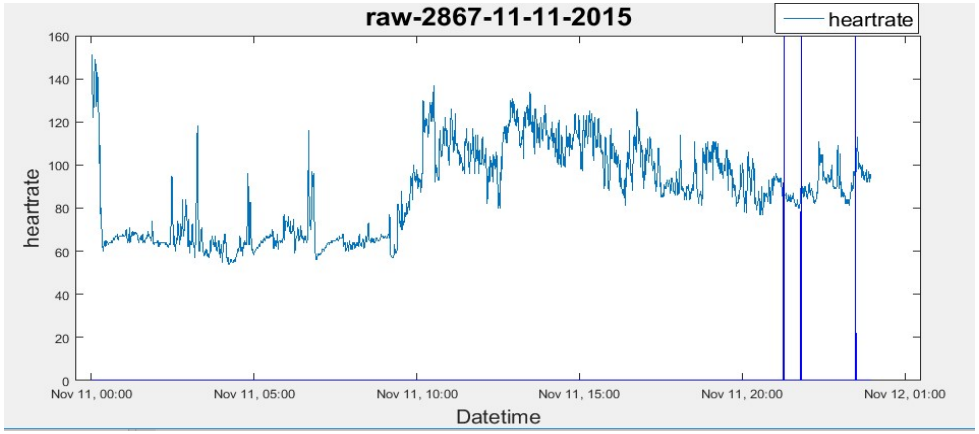
	1 Patient	2 Type	3 Type1	4 StartTS	5 EndTS	6 AD	7 SLS	8 SLR
1	212	2	'ID'	'12/17/15 23:12'	'12/17/15 23:17'	'2'	''	''
2	212	2	'ID'	'12/18/15 19:24'	'12/18/15 19:27'	'2'	''	''
3	212	5	'RS2'	'12/20/15 18:37'	'12/20/15 18:39'	'1'	'1'	''
4	212	5	'RS3'	'12/21/15 20:52'	'12/21/15 20:54'	'1'	'1'	''
5	212	5	'RS1'	'12/25/15 13:27'	'12/25/15 13:29'	'2'	'1'	''
6	212	6	'DF'	'12/25/15 14:34'	'12/25/15 14:37'	'1'	''	'1'
7	212	5	'RS1'	'12/27/15 13:38'	'12/27/15 13:40'	'3'	'1'	''
8	212	5	'RS2'	'12/31/15 15:46'	'12/31/15 15:50'	'2'	'1'	''
9	212	5	'RS3'	'12/31/15 19:42'	'12/31/15 19:44'	'1'	'1'	''
10	212	5	'RS2'	'1/1/16 18:01'	'1/1/16 18:03'	'2'	'1'	''

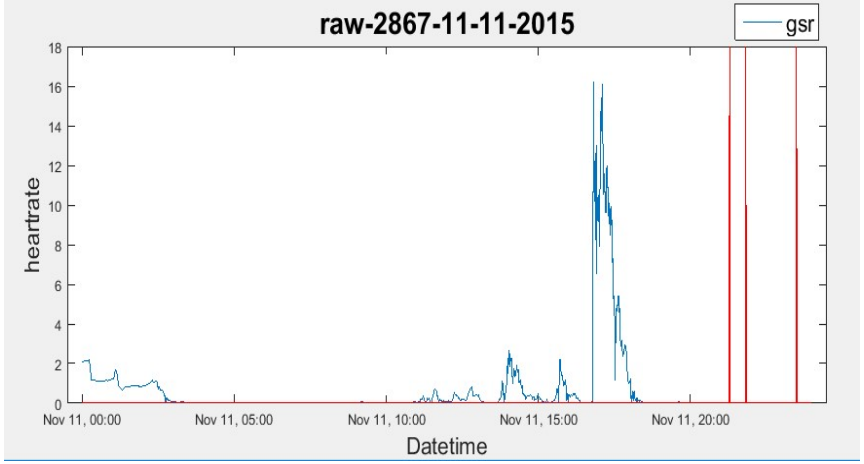
Sensor Data Example

1	patient	datetime	skin_temp	heartrate	gsr
2	2867	10/27/2015 10:41	77	71	5.83E-05
3	2867	10/27/2015 10:42	80.6	69	6.40E-05
4	2867	10/27/2015 10:43	82.4	68	6.29E-05
5	2867	10/27/2015 10:44	83.3	67	6.26E-05
6	2867	10/27/2015 10:45	84.2	68	6.33E-05
7	2867	10/27/2015 10:46	85.1	70	6.23E-05
8	2867	10/27/2015 10:47	85.1	70	6.33E-05
9	2867	10/27/2015 10:48	86	69	6.47E-05
10	2867	10/27/2015 10:49	86	68	6.54E-05
11	2867	10/27/2015 10:50	86	69	6.58E-05
12	2867	10/27/2015 10:51	86.9	75	7.07E-05
13	2867	10/27/2015 10:52	86.9	95	6.95E-05
14	2867	10/27/2015 10:53	86.9	91	6.84E-05

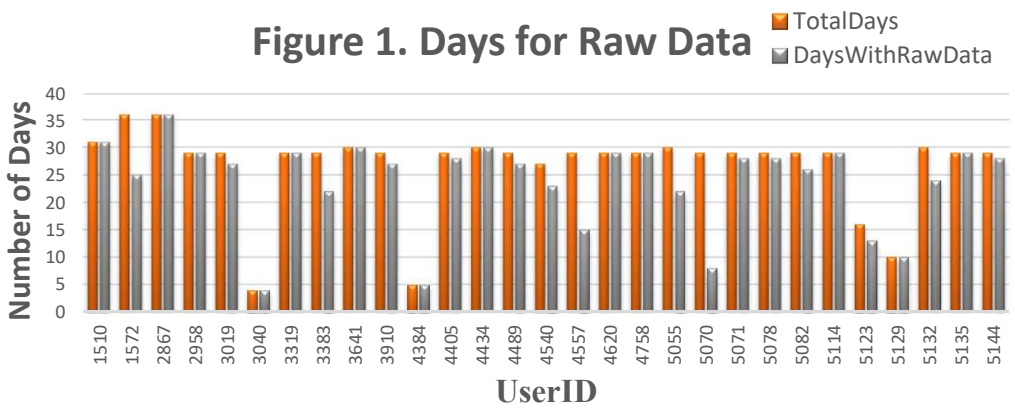
2. Data Visualization

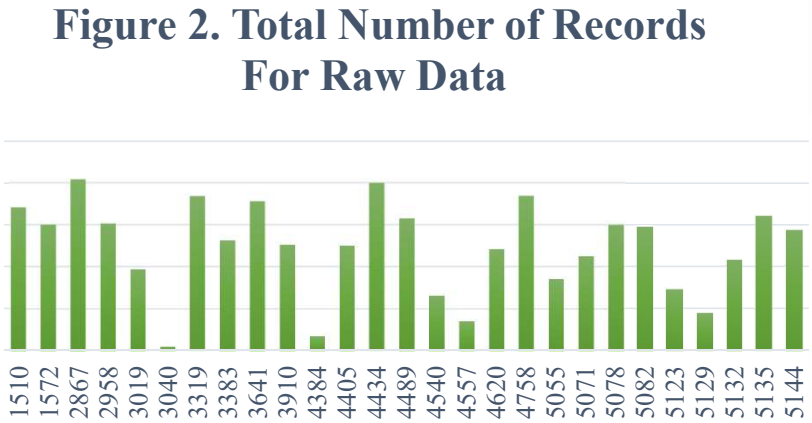


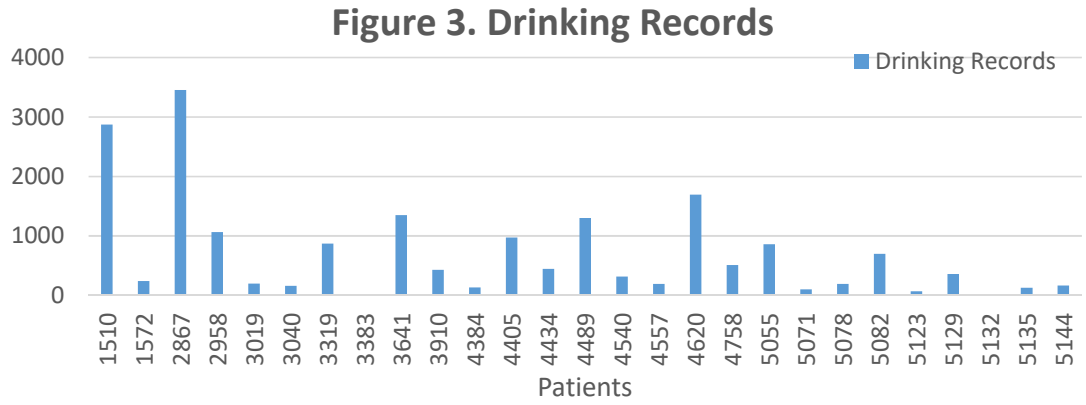




3. Data Statistics





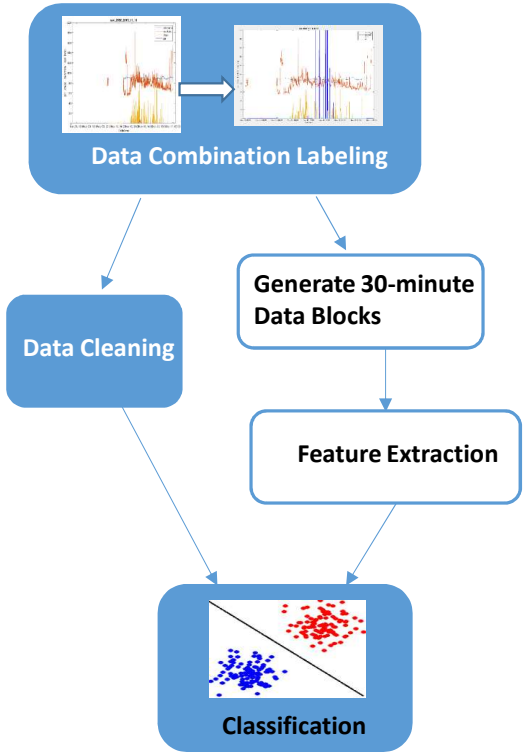


Contents

Data Analysis Methods Overview

Method 1: Drinking record prediction Pipeline

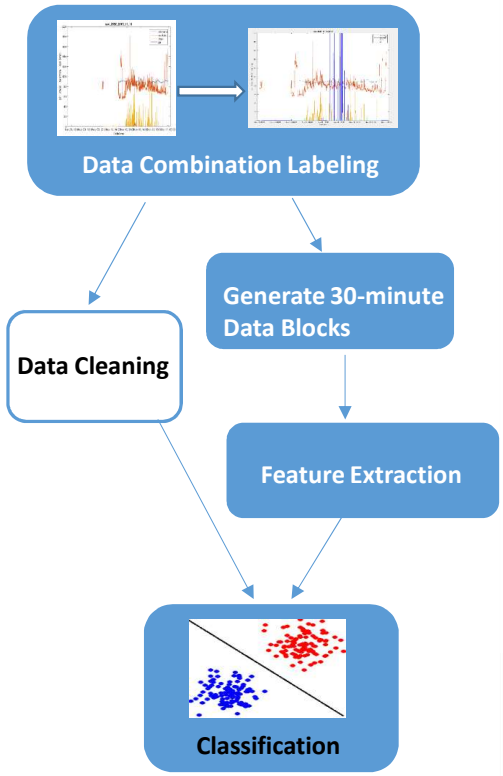
1. Data combination and labeling
2. Data cleaning:
   1) Gaps and insufficient data removal
   2) Smoothing and outliers removal
3. Classification



Data Analysis Methods Overview

Method 2: Drinking episode prediction statistical pipeline

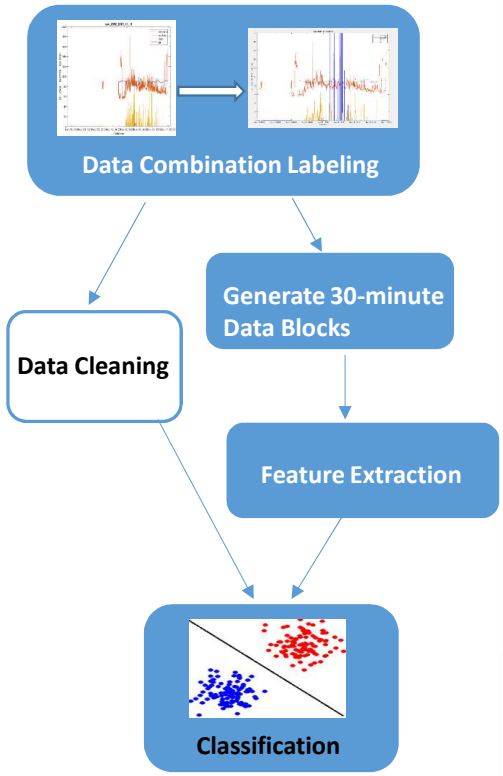
1. Data Combination and Labeling
2. Generate 30-minute data blocks
3. Extract statistical features from 30-minute data blocks
4. Principal component analysis
5. Classification



Data Analysis Methods Overview

Method 3: Drinking episode prediction deep learning pipeline

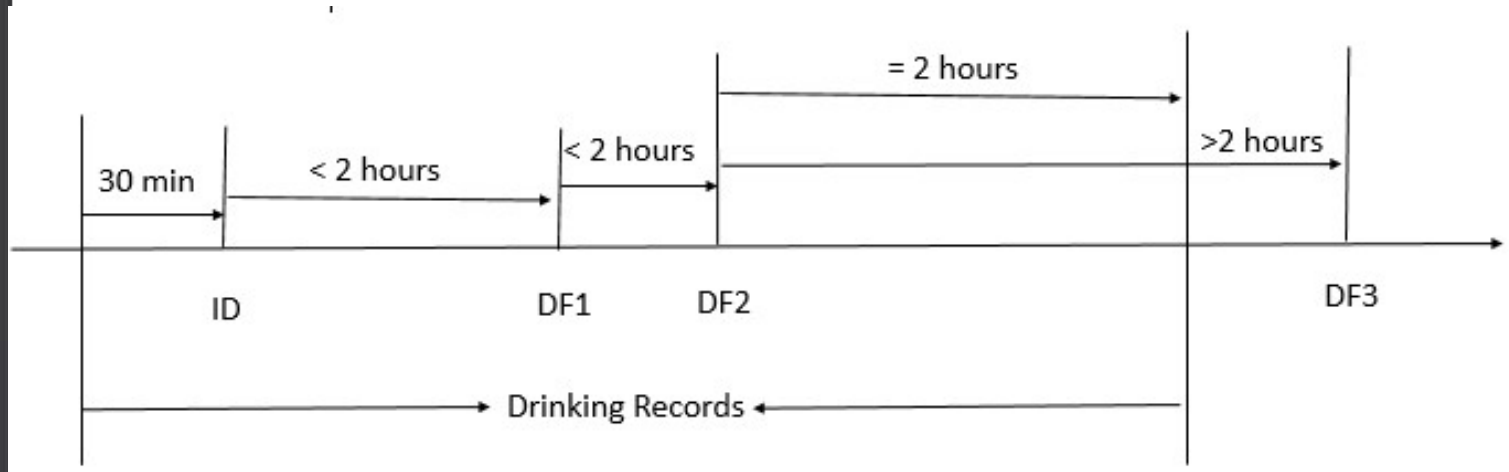
1. Data Combination and Labeling
2. Generate 30-minute data blocks
3. Convert 30-minute data blocks into spectrogram
4. Extract deep learning features from spectrogram
5. Classification



Contents

1. Data Combination and Labeling

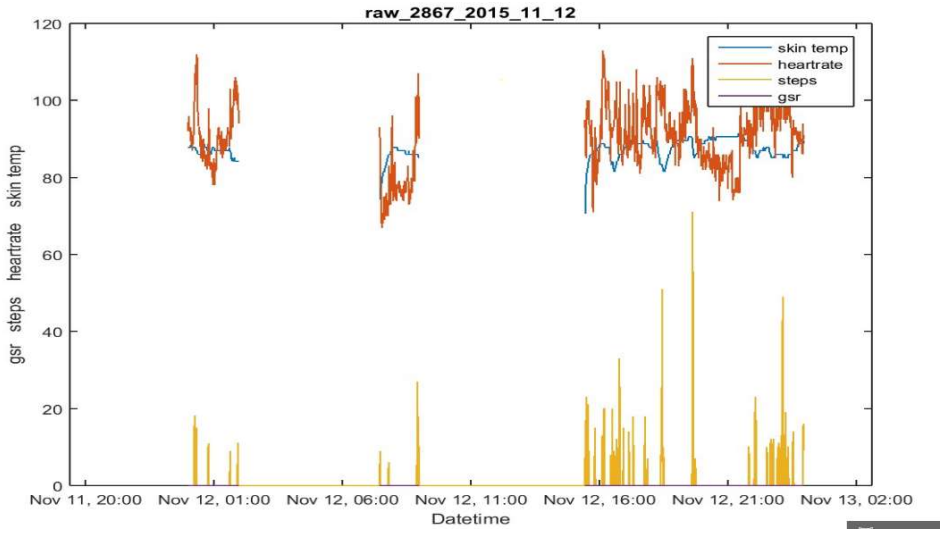
1. Combine raw sensor data with survey data
2. Find initial drinking and drinking follow-ups that have a time difference less than 2 hours with its previous drinking behavior
3. Label data points that fall into [ID - 30 minutes, Last DF + 2 hours] as drinking



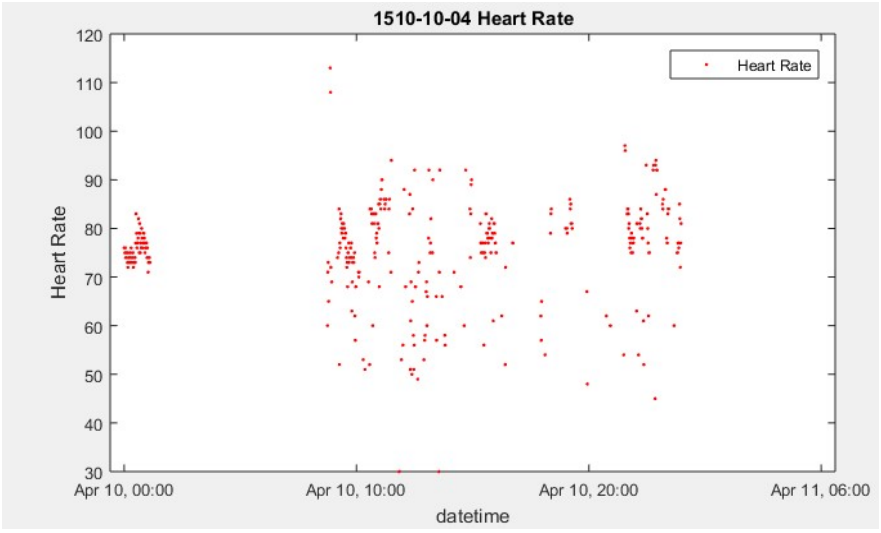
ID: Initial drinking
DF1: Drinking follow-up 1
DF2: Drinking follow-up 2
DF3: Drinking follow-up 3

2. Data Cleaning Step 1: Gaps and Insufficient Data Removal

1) Gaps: There is no data within 10-minute window
2) Insufficient Data: Less than 5 data points within 10-minute window



Example for Gaps

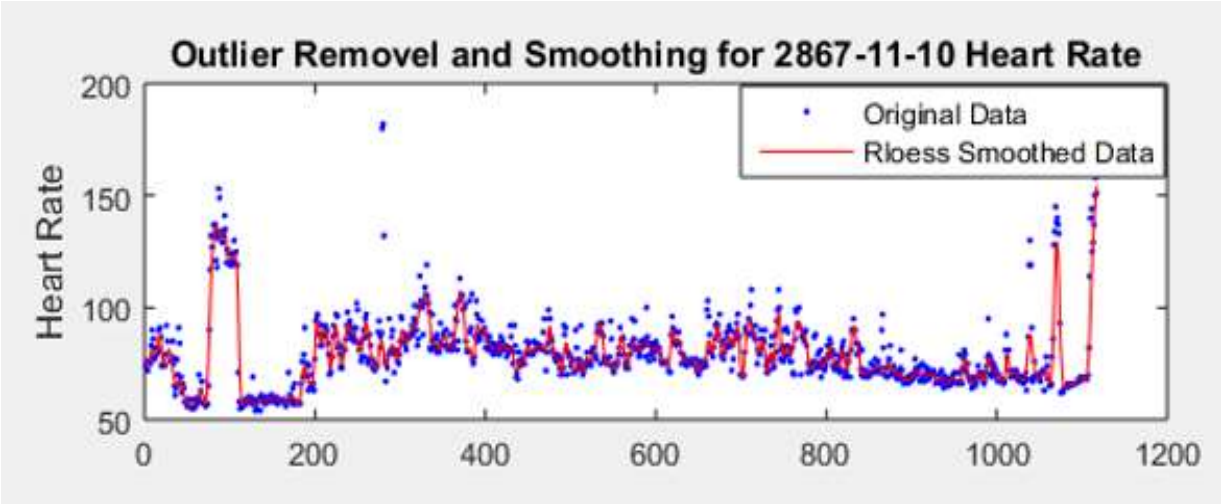


Example for Insufficient Data

2. Data Cleaning Step 2: Smoothing and Outliers Removal

Use Lowess to smooth the data and remove outliers

1) Window Size: 1% of the data
2) Outliers: Two standard deviations away from the fitted curve



Classification

Four Classifiers:

1) Naïve Bayes
2) Bayes Network
3) Logistic Regression
4) J48 Decision Tree

Contents

2. Generate 30-Minute Data Blocks

Input: Labeled one-dimensional signal

Requirement:

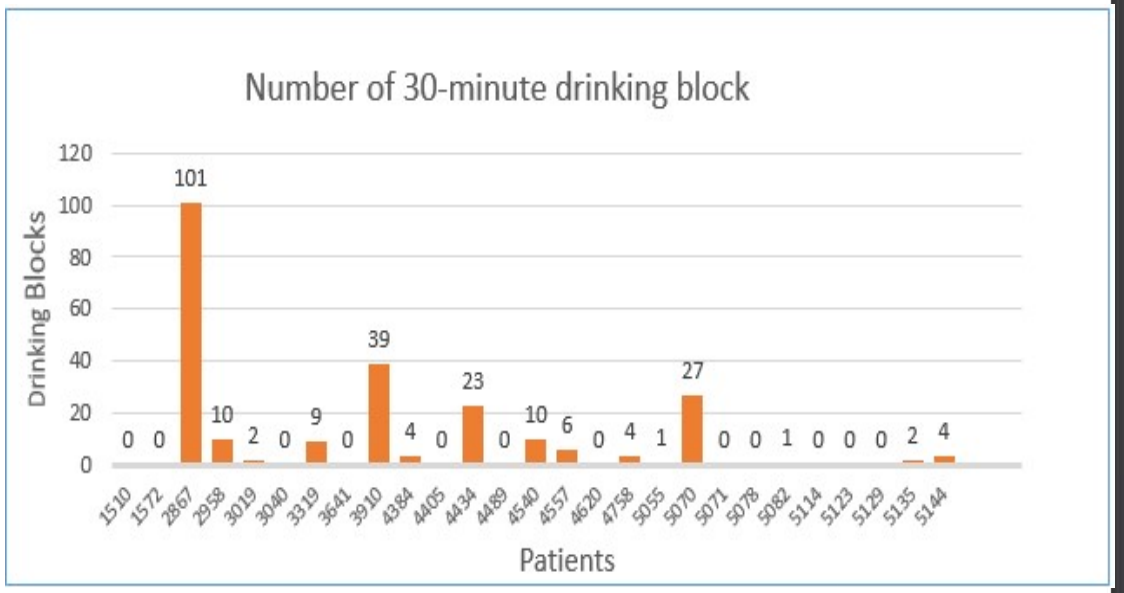
1)There is no missing value in 30-minute window

2) All the data points in the 30-minute window are labeled as the same type

Output:

1) positive data block: if all 30 data points are drinking

2) negative data block: if all 30 data points are non-drinking



3. Statistical Feature Extraction

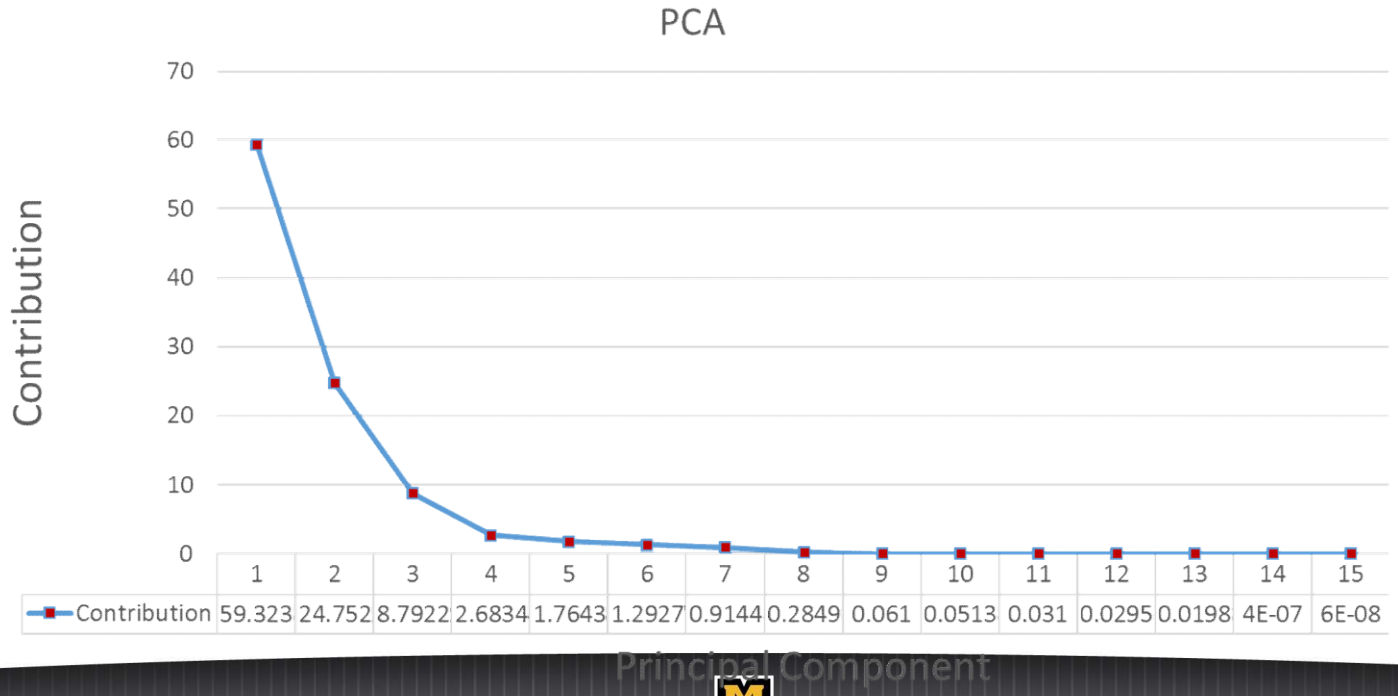
Statistical Features:

- Mean: $\frac{1}{n}\sum_{i=1}^{n} x_i$
- Standard Deviation: $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2}$
- Skewness: $\frac{E(x-\mu)^3}{\sigma^3}$
- Slope: The slop of linear regression fitted on the data block
- Coefficient of Variance: Std/Mean (measure spread relative to mean)

4. Principal Component Analysis

Rule: Contribution larger than 0.1 percent

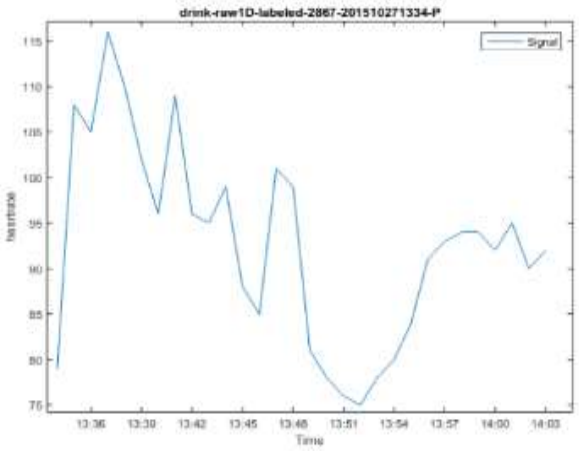
Result: 8 principal components were chose

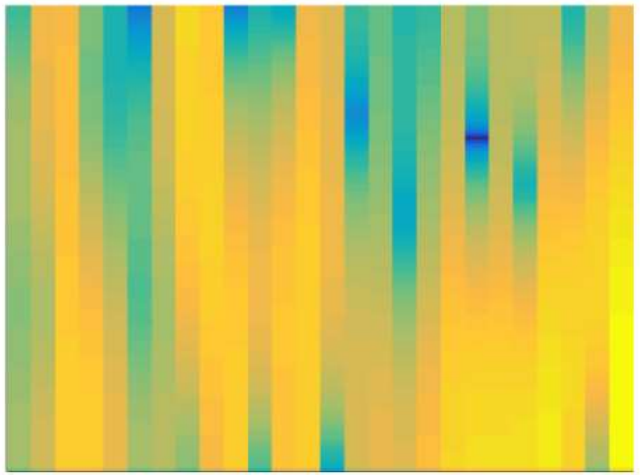


Contents

3. Convert 30-minute data block into Spectrogram

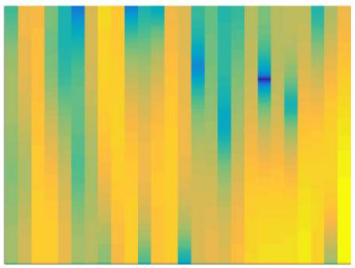
- Window size: 5
- Overlap: window size – 1
- Sample rate: 1 minute
- Normalized
- Color





4. Generate Cifar 10 Features from Spectrogram

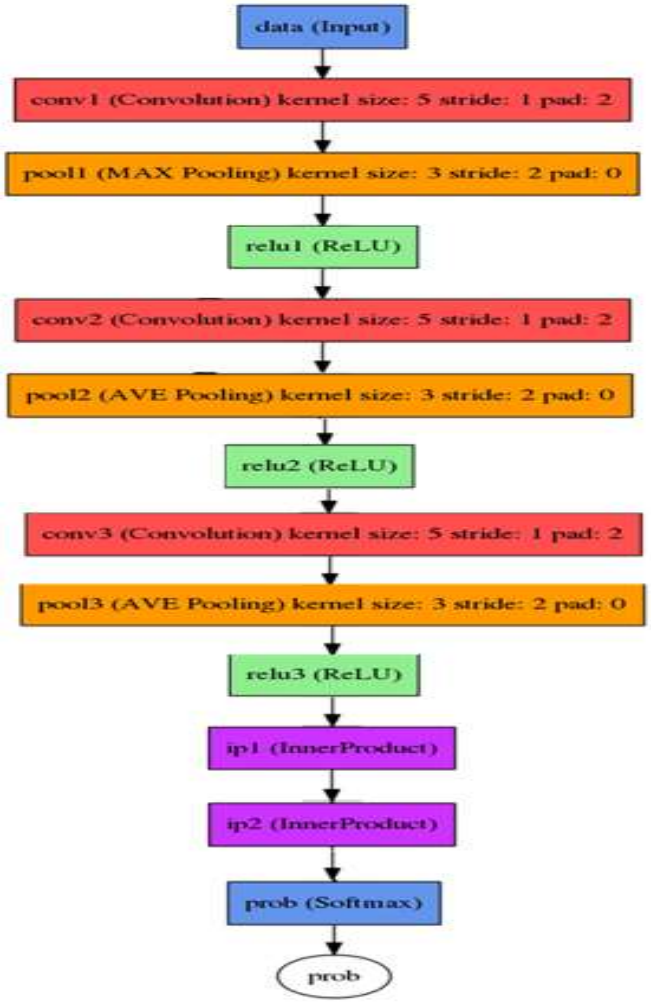
Use pre-trained model to do classification on Spectrogram to generate 10 probabilities for each Cifar 10 category



Spectrogram

feature1	feature2	feature3	feature4	feature5	feature6	feature7	feature8	feature9	feature10
0.089072	0.002002	0.859683	0.012233	0.018286	0.001166	0.009215	0.001173	0.000807	0.006364
0.046998	0.000325	0.939184	0.001438	0.006561	0.00014	0.003386	0.000275	0.000211	0.001482
0.038424	0.000116	0.933268	0.013674	0.003716	0.000372	0.008876	0.000153	0.000325	0.001077
0.020509	0.00027	0.971109	0.001356	0.004789	7.52E-05	0.00105	0.000193	0.000101	0.000549
0.029968	0.000247	0.952733	0.00348	0.010214	0.000143	0.001396	0.000361	0.000176	0.001283
0.010048	0.000139	0.977345	0.004124	0.00583	0.000179	0.001363	0.000235	0.000104	0.000633
0.024207	0.000104	0.96901	0.001258	0.003482	5.41E-05	0.001443	0.000126	7.06E-05	0.000246
0.030022	0.000789	0.934306	0.00397	0.024449	0.000206	0.004072	0.000759	0.000185	0.001243

Cifar 10 Features

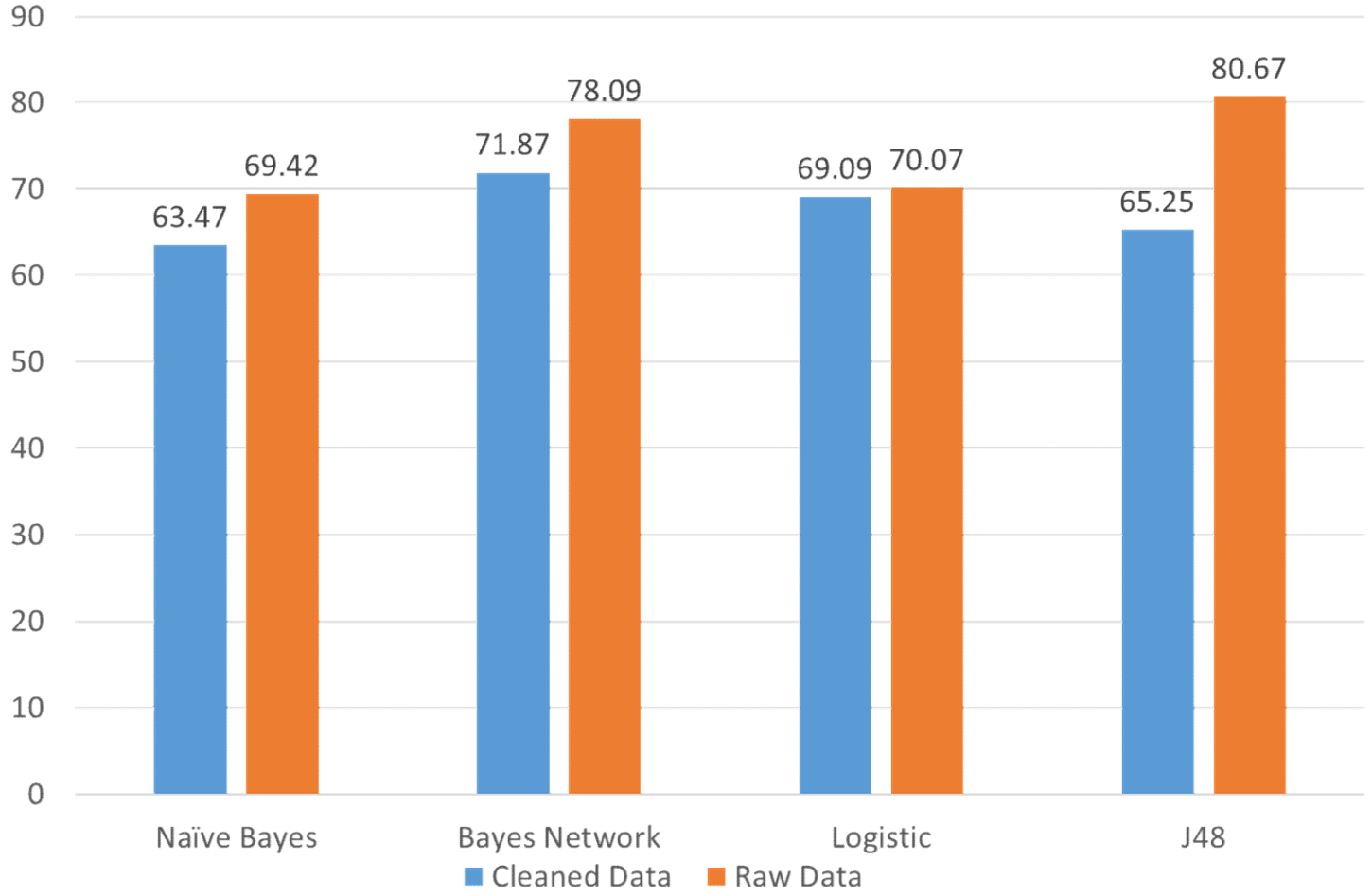


Contents

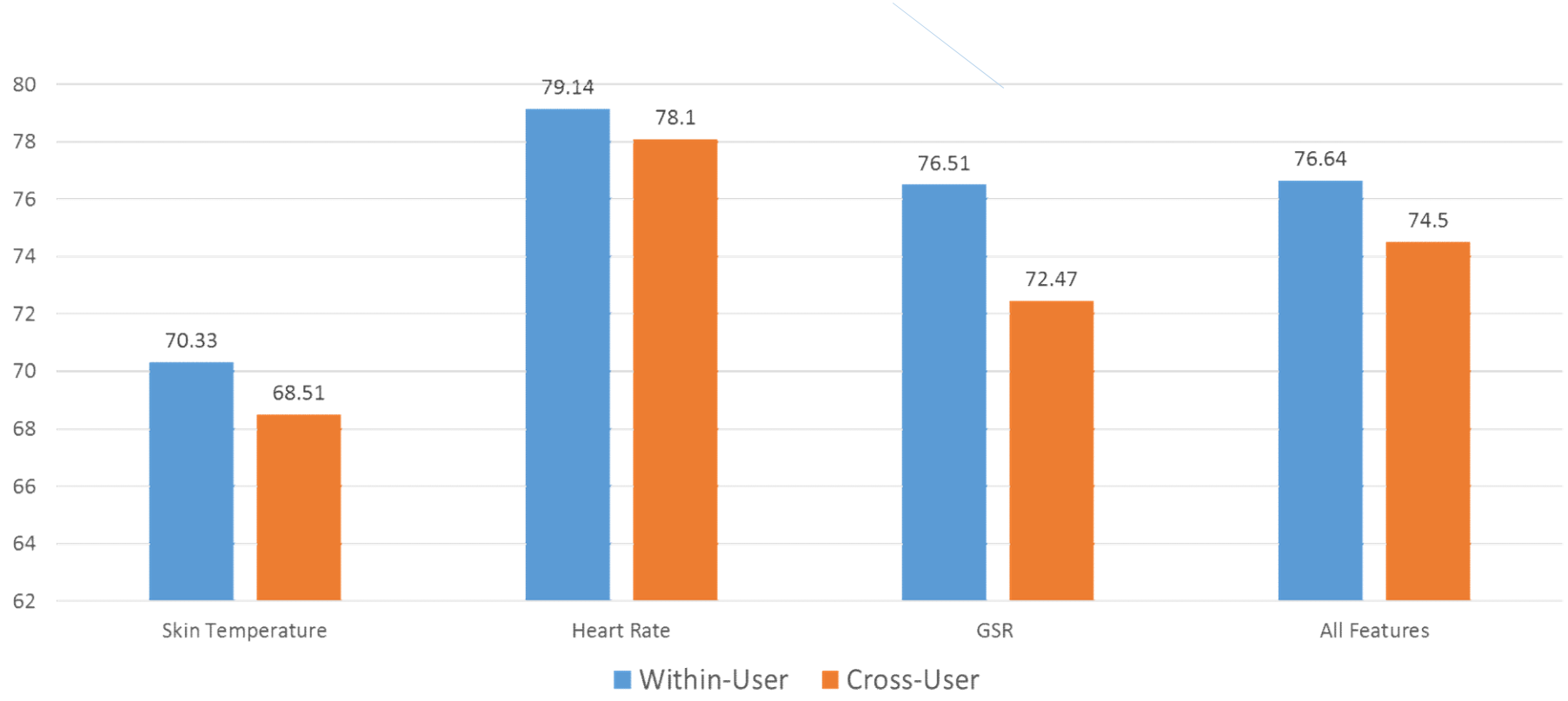
Training and Testing Dataset

1. Dataset for method 1: drinking records prediction pipeline
    1) All users
    2) Data type: 1-minute record
    3) 14856 drinking and 14856 non-drinking
    4) 66% for training, 34% for testing
2. Dataset for method 2 and method 3: drinking episode prediction
    1) Three users: 2867, 3641, 5055
    2) Data type: 30-minute data blocks
    3) User 2867: 101 drinking and 101 non-drinking
    4) User 3641: 36 drinking and 36 non-drinking
    5) User 5055: 26 drinking and 26 non-drinking
    6) 66% for training, 34% for testing

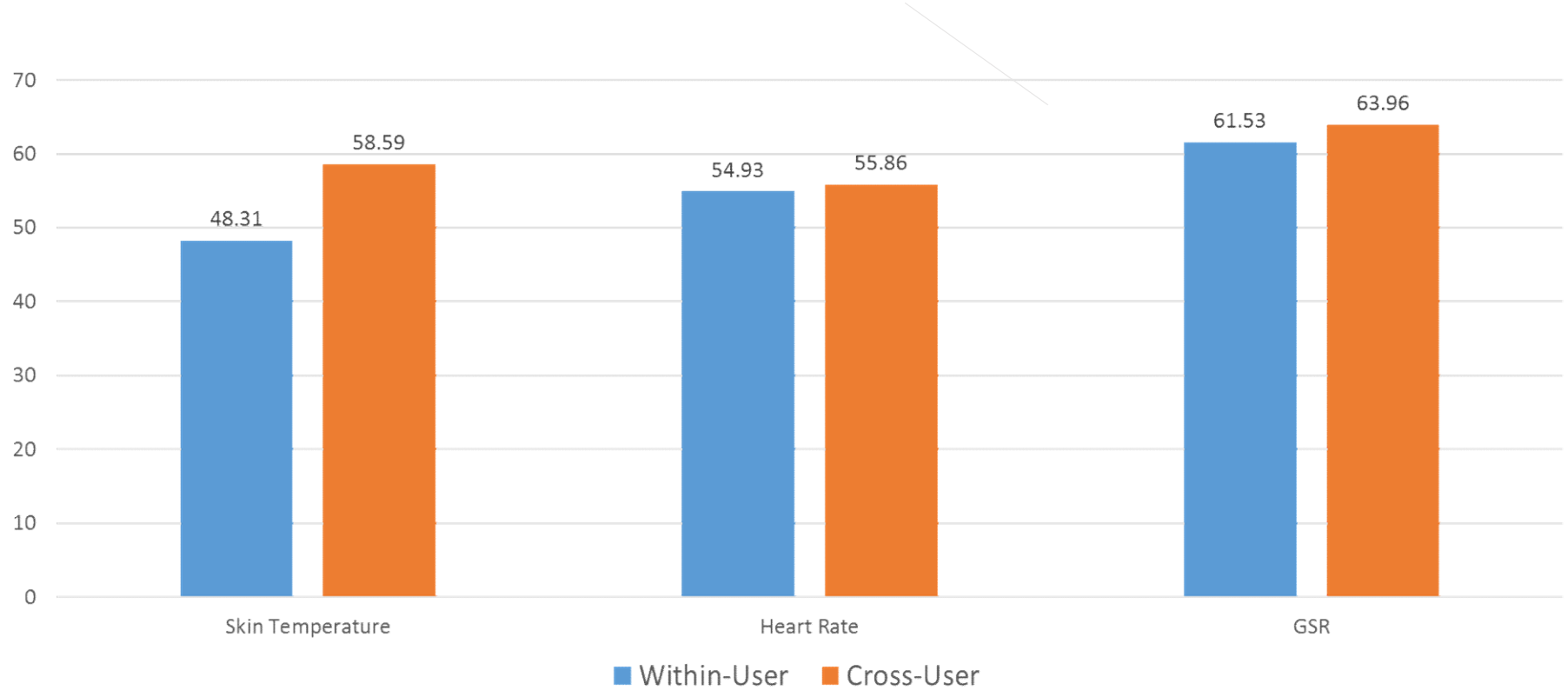
1. Result for Drinking Record Pipeline



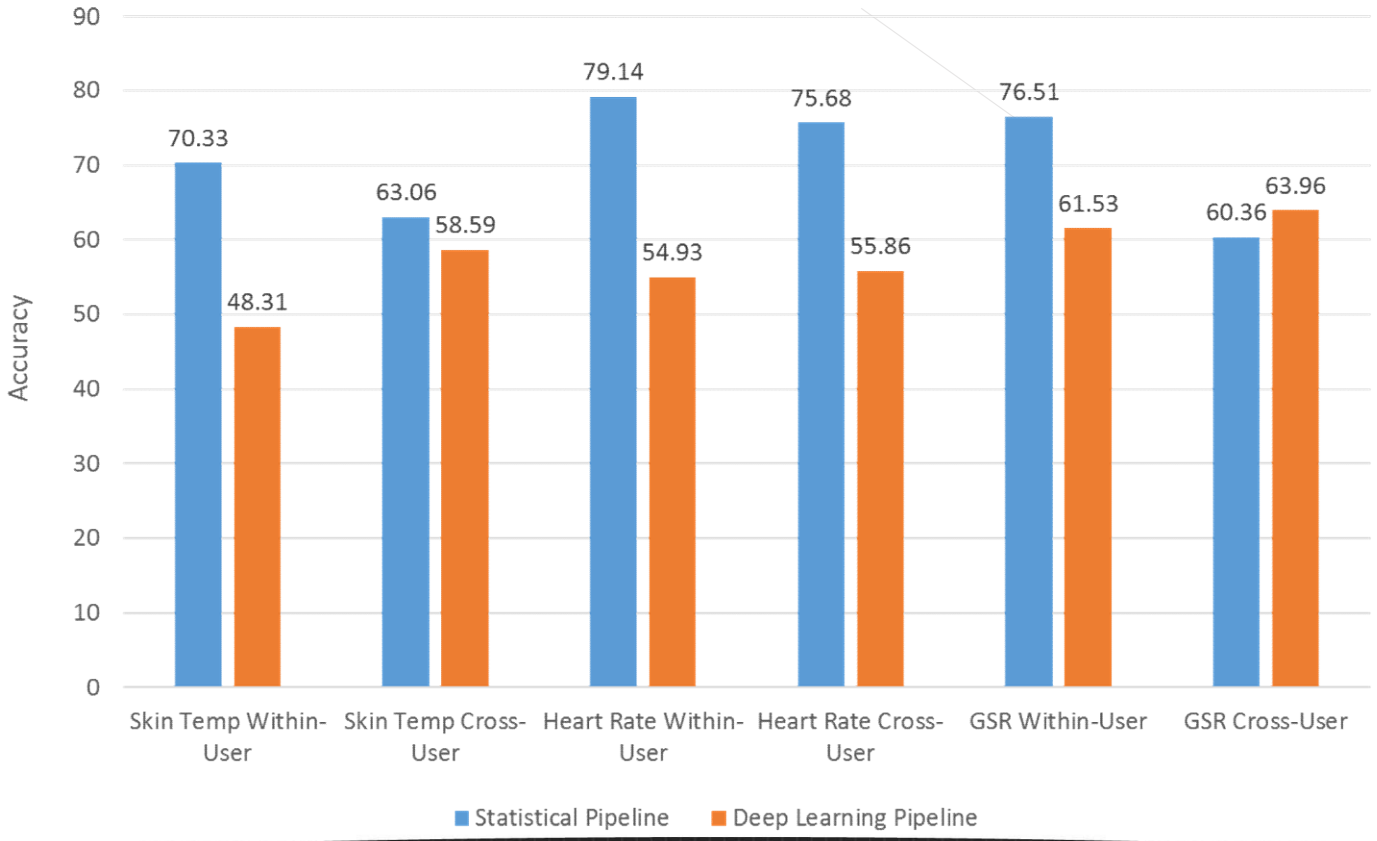
2. Result for Drinking Episode Statistical Pipeline



3. Results for Drinking Episode Deep Learning Pipeline



4. Statistical Pipeline VS Deep Learning Pipeline



Contents

- Data Analysis Methods
  1. Data Analysis Methods Overview
  2. Method 1: Drinking Record Prediction Pipeline
  3. Method 2: Drinking Episode Prediction Statistical Pipeline
  4. Method 3: Drinking Episode Prediction Deep Learning Pipeline
- Experiment Results and Comparison
- Conclusion and Future Work

Conclusion and Future Work

- Conclusion
  - Raw data has better result than cleaned data on drinking record prediction
  - Within-user result is much better than cross-user result
  - Statistical pipeline has better result than deep learning pipeline
  - Heart rate is the most significant feature in drinking episode prediction
- Future work
  - Take alcohol amount into account for labeling
  - Try more deep learning models
  - Apply the methods in this thesis to other larger amount of data


Thank You!
Question?
University of Missouri