PHYSIOLOGICAL DATA ANALYSIS – ALCOHOL DRINKING PREDICTION USING STATISTICAL AND DEEP LEARNING METHODS

A Thesis

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

CAN LI

Dr. Yi Shang, Thesis Advisor

MAY 2017

The undersigned, appointed by the dean of the Graduate School, have examined the thesis entitled

PHYSIOLOGICAL DATA ANALYSIS – ALCOHOL DRINKING PREDICTION USING STATISTICAL AND DEEP LEARNING METHODS

presented by Can Li,

a candidate for the degree of master of science,

and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Yi Shang

Dr. Timothy Trull

Dr. Yunxin Zhao

ACKNOWLEDGEMENTS

I would like to thank my academic advisor, Dr. Yi Shang for not only helping me finish my thesis but also helping me build my career. In the process of completing my thesis, he is so patient for showing me the method for how to do a good research. It is him that gave me the opportunity to do great projects with him, which is very useful to my career.

I would also like to thank my thesis committee members, Dr. Tim Trull and Dr. Yunxin Zhao. Thank them for spending time on reviewing my paper, giving their valuable suggestions, and attending my defense.

Moreover, I would like to give my thanks to Jodette Lenser for all her helps that she offered me during the process of achieving my academic degree. Especially thank her for submitting M2 form for me when I was in Saint Louis.

Finally, I would like to thank all the team members in Dr. Shang’s lab, especially Peng, Guang, and Nick. Thank them for sharing their valuable knowledge and suggestions to me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Alcohol craving can cause many problems for people's life. However, there are very few related works doing alcohol prediction based on physiological data, except some from our lab. The goal of this research is to predict whether people have alcohol drinking or not from real physiological data in order to help them with drinking problems. The raw physiological data in this work include skin temperature, heart rate, galvanic skin response (GSR), steps, and calories. The data was collected from 29 users with basic watch and the reading frequency is one record per minute.

In this thesis, three data analysis pipelines, drinking record prediction pipeline, drinking episode statistical pipeline, and drinking episode deep learning pipeline, are implemented. The drinking record prediction pipeline is doing prediction based on one-minute record. The drinking episode pipeline is doing prediction based on thirty-minute episode. Statistical features are extracted from the thirty-minute data blocks. The drinking episode deep learning pipeline is doing prediction based on thirty-minute episode as well. In this deep learning pipeline, one dimensional signal is converted to spectrum graph. Then use Cifar 10 model to extract deep learning features from the spectrum graph. After that apply machine learning methods on the deep learning features to do the classification.

Within-user and cross user experiments are conducted in this thesis because different users may have different reaction to alcohol. Different models are found for different users and general model is discovered for cross-users. Balanced data is used for training and testing, so the baseline accuracy is 50%. The accuracy for within-user is up to 88.89% and the accuracy for cross-user is 75.68%, which indicates that the within-user result is much better than cross-user result. In order to find the most significant feature in alcohol drinking prediction, experiments are also conducted on skin temperature only features, heart rate only features, and GSR only features. The results show that heart rate contributes most in the alcohol drinking prediction.

1. INTRODUCTION

Nowadays data is the most valuable resources. Whoever has data resources has the greatest opportunities. Devices with wearable sensors are becoming more and more popular. For example, Apple has iwatch, which can collect your physiological data, heart rate, skin temperature and so on. And iHealth has a product called iHealth Align, which can track your glucose level.

Only having data is not enough, it is better to know how to use the data. Artificial Intelligence is the hottest discipline right now, which uses machine learning to discover the underlying information from available data and predicts what happens next. For example, with physiological data, an app can be created to use machine learning methods to predict what kind of diseases people may have in the future, which can help them prevent these diseases happening.

The project is a part of a big project called mobile ambulatory assessment system [1] [2] [3]. Our lab and psychology department cooperated on this big project. Psychology department wants to discover how alcohol drinking could affect people's behavior. They found some patients who have alcohol drinking problem. These patients would wear basic watch or sensor suits and have an android phone with an application. This application was developed by other members in our lab. Patients can use this app to do scheduled survey. They could report the time when they had drinking and other information. This app also has some random pop-up surveys. The survey data and the sensor data will be sent to the server [1] [2] [3]. This is how the data was collected.

Besides the data collection part, the mobile ambulatory assessment system includes the data analysis part that is what this thesis about. The data used in this work is from the basic watch. In this thesis three different pipelines are going to be implemented to predict alcohol drinking. The first pipeline is dinking record pipeline, which will do the prediction based on one-minute record. The second pipeline is drinking episode statistical pipeline. In this pipeline, statistical features, like mean, standard deviation and so on, are going to be extracted from the raw data. Then traditional machine learning methods, like Naïve Bayes, Bayes Network, Logistic Regression, and J48 Decision Tree, will be applied to the extracted features to get the prediction results. Unlike the second pipeline using statistical features extracted from the raw data, the third pipeline will use deep learning method, Cifar 10, to extract deep learning features. Convert raw data into spectrum graph, and then using Cifar 10 to extract deep learning features from the spectrum graph. Then run machine learning methods on the deep learning features to get the prediction results.

Currently the mobile ambulatory assessment system is not completely automatic. The data collection part and the data analysis part are two separate parts. In the future, after the data analysis part is completely done, the data collection part and data analysis part can be integrated together into an app to make the whole mobile ambulatory assessment system run automatically. Then the app can predict when people have an alcohol drinking, which can help people who have drinking problems.

1.1 Problems and Motivation

Large amount of time has been spent in implementing the mobile ambulatory assessment system and a lot of physiological data has been collected so far. The goal of this project is to find out how alcohol drinking can affect human's behavior. Data analysis needs to be done on the physiological data in order to find a good model to predict when people have alcohol drinking.

There are some existing works that try to analyze human's behavior based on physiological data. For example, there is one paper about how smoking affects human's body by analyzing physiological data. But there is not too much work on alcohol drinking prediction. So this is a pretty new research area and it is very worthy to do research on. Although there have been some previous work on alcohol drinking prediction on other different physiological data in the lab, these are some preliminary work. They do not use the most appropriate methods to deal with data or have not found the good model for alcohol drinking prediction. For example, their physiological data is collected by every 5 seconds. They label the data based on 5-second record and do the drinking prediction. The problem here is that alcohol can have a long time effect on our body; usually the duration of the effect is several hours. If using the 5-second record to do the prediction, the results will not be accurate because there are too much over lapping information between the records that are close to each other on the time stamp. Drinking episode prediction is used to solve this problem.

1.2 Contributions

The main contributions are from three aspects, including how to label data, feature extraction, multiple new pipelines, and machine learning methods.

1 Data labeling

As mentioned in the problem and motivation section, previous work label the data based on 5-second reading. Those records that are close to each other in time stamp contain too much overlapping information, which results in inaccurate and unreasonable prediction.

In this work, 30 minutes data blocks are generated and labeled. In this way there is no more overlapping information between any two data blocks.

2 Feature extraction

Besides extracting statistical features, like mean, standard deviation, from the raw data, deep learning features are extracted. The general process is that transform the raw data into spectrum graph first, then use deep learning model, Cifar 10 to extract deep learning features from the spectrum graphs.

3 Multiple new pipelines

Three pipelines are implemented in this work. The first one is drinking record prediction pipeline, doing prediction based on one-minute record. The second one is drinking episode statistical pipeline, doing prediction based on 30-minute data blocks.

The main process of the second pipeline is that extract statistical features from raw data, and then apply machine learning methods. The third pipeline is using deep learning model to extract deep learning features from spectrum graphs and then apply machine learning methods. Because there are 29 users in this research and each patient has different habits and physiological data, different model is appropriate on different situations.

4 Machine learning methods

Because the physiological data we collect is very noisy, statistical features and traditional machine learning methods can’t produce good prediction model. Deep learning methods are applied to extract underlying physical information in order to find good prediction models.

1.3 Thesis Outline

The thesis is structured as following:

Chapter 1 is about introduction to this project and what are the problems and motivations for doing this project.

Chapter 2 introduces the background of this project and some related work.

Chapter 3 is about data overview, data preprocessing and data cleaning.

Chapter 4 talks about the implementation of multiple new pipelines.

Chapter 5 is about experimental design, experimental results, and results analysis.

Chapter 6 is about future work.

Chapter 7 is the conclusion of the whole work.

Chapter 8 is the references.

2. Background and Related Work

2.1 Background

This chapter introduces the background about some machine learning methods because this thesis work uses a lot of machine learning knowledge. Some very common supervised learning classification methods, such as Naive Bayes, Bayes Network, Logistic Regression and J48 Decision Tree, are used to generate the experimental results. Statistical methods are used to generate statistical features in the feature extraction phase. But statistical features may not be deep enough to explain all the underlying information in the experimental data. So deep learning method is used to extract deep learning features.

Cifar 10 dataset were collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton [6]. The whole dataset consists of 60000 32x32 color images from 10 classes. And each class has 6000 images. 5000 of them are training dataset and 1000 of them are testing dataset [6]. The following is a Cifar 10 example of 10 random images from each class:

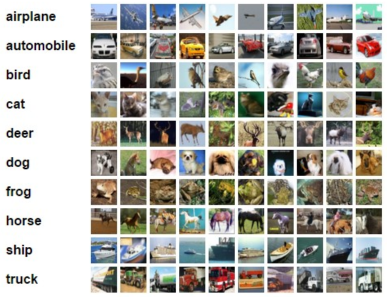


Figure 1. Cifar 10 Dataset Example

The goal is to use convolutional neural network and Cifar 10 dataset [6] [7] to extract deep learning features from the spectrum graphs of the raw data.

Although traditional multilayer perceptron (MLP) models were successfully used for image recognition, they don’t scale well to higher resolution images due to the full connectivity between nodes. Convolutional neural networks (CNN) are biologically inspired variants of multilayer perceptron, designed to emulate the behavior of a visual cortex [8]. These models mitigate the challenges posed by the MLP architecture by exploiting the strong spatially local correlation present in natural images. So CNN works very well for higher resolution images. CNN consists of multiple layers of receptive fields. These are small neuron collections which process portions of the input image. The outputs of these collections are then tiled so that their input regions overlap, to obtain a higher-resolution representation of the original image [9] [10] [11].

2.2 Related Work

Although there are too many papers that are doing drinking prediction with physiological data, there are still many papers that are doing research about human activities prediction from physiological data. Moreover there is one paper from the lab doing drinking prediction and the other paper doing mood dysregulation prediction.

Paper [5] talks about how cocaine affects human’s body. In their study, they calculate a lot of ECG related features, which is very helpful to this research because there is also heart rate feature in this study. This related work let their users to take different dosage of cocaine to see how their body reacts. In my own study, different users may have different reaction to alcohol and the same user may drinks different amount of alcohol each time. So it may not be able to find a general model for all the users. Instead, it is better to find different model for different users. This paper also tries to find the user’s recovery time from cocaine intake. The alcohol also has long effects on human’s inner body features. So it is better to do prediction based on time period, not only based on each record because our data is collected every minute. Paper [12] is about cocaine detection on heart rate features. Paper [13] is doing research on the effect of cocaine use on heart.

Paper [3] is doing drinking prediction based on physiological data collected from wearable sensors. This paper uses ECG and respiration features. Statistical features are extracted from every one-minute window size. A very good data cleaning process is used. The drinking prediction is done based on every minute. For the experimental design, [2] uses balanced data, which results a 50% baseline.

Paper [4] is predicting mood dysregulation from physiological data. The raw data this paper is using is also collected from wearable sensors. This paper also has a very good data cleaning pipeline. Both within-user and between-user experimental design are conducted in this paper. The prediction in this work is based on every 5-second record and it has a very high accuracy.

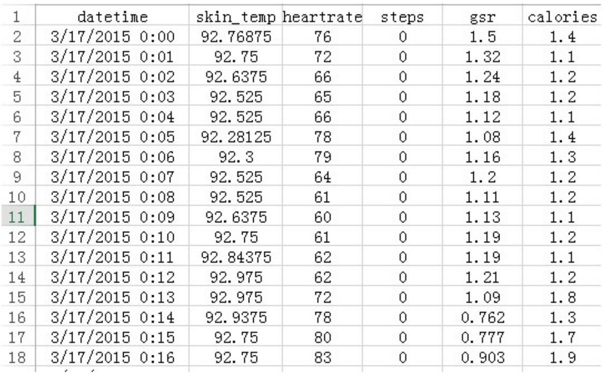
3. Data Preprocessing and Cleaning

This chapter talks about data preprocessing and cleaning. Because it is a common and big part of the three pipelines implementation, so it is put in a separate chapter.

3.1 Data Overview

As introduced in the background section, the data in this thesis is collected from the mobile ambulatory assessment system. The data is composed of two kinds of data. One is the physiological data, and it will be called raw data later. And the other is survey data.

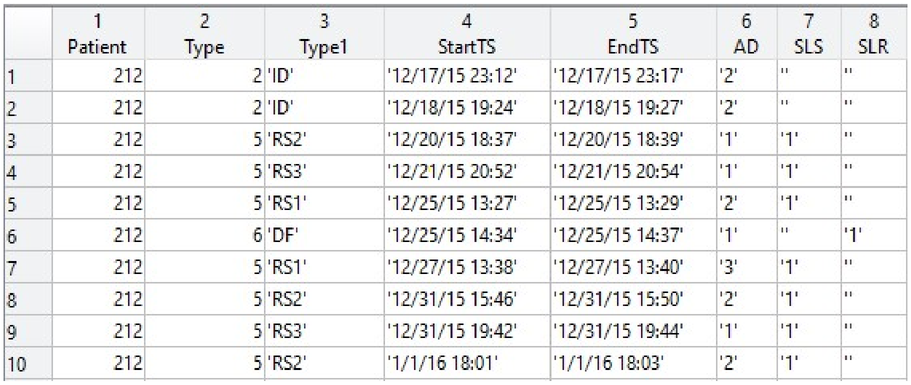
The raw data is collected by every minute. It consists of five features. They are skin temperature, heart rate, steps, galvanic skin response (use GSR later), and calories. Also the raw data has a time stamp. Later the time stamp can be used to combine with the survey data. Figure 2 is a screen shot of the raw data.



1	datetime	skin_temp	heartrate	steps	gsr	calories
2	3/17/2015 0:00	92.76875	76	0	1.5	1.4
3	3/17/2015 0:01	92.75	72	0	1.32	1.1
4	3/17/2015 0:02	92.6375	66	0	1.24	1.2
5	3/17/2015 0:03	92.525	65	0	1.18	1.2
6	3/17/2015 0:04	92.525	66	0	1.12	1.1
7	3/17/2015 0:05	92.28125	78	0	1.08	1.4
8	3/17/2015 0:06	92.3	79	0	1.16	1.3
9	3/17/2015 0:07	92.525	64	0	1.2	1.2
10	3/17/2015 0:08	92.525	61	0	1.11	1.2
11	3/17/2015 0:09	92.6375	60	0	1.13	1.1
12	3/17/2015 0:10	92.75	61	0	1.19	1.2
13	3/17/2015 0:11	92.84375	62	0	1.19	1.1
14	3/17/2015 0:12	92.975	62	0	1.21	1.2
15	3/17/2015 0:13	92.975	72	0	1.09	1.8
16	3/17/2015 0:14	92.9375	78	0	0.762	1.3
17	3/17/2015 0:15	92.75	80	0	0.777	1.7
18	3/17/2015 0:16	92.75	83	0	0.903	1.9

Figure 2. Raw Data Example

The survey data contains totally 112 columns. It has more than 100 survey questions. The useful columns are extracted from the survey data. Figure 3 is screen shot for survey data.



	1	2	3	4	5	6	7	8
	Patient	Type	Type1	StartTS	EndTS	AD	SLS	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'	''

Figure 3. Survey Data Example

Take a look at Figure 3, the survey data contains information like survey start time and end time. Column 6 is a survey question that: how much alcohol did you drink. For simplicity, I use ‘AD’ to present it. Survey question in column 7 is how much alcohol did you drink since last survey and I use ‘SLS’ for short. Column 8 is also a survey question. The question is how many alcohol you drank since last random survey. The survey start time, end time and these three survey questions can be used to label the raw data.

3.2 Data preprocessing

3.2.1 Survey Data preprocessing

Survey data preprocessing includes extract useful survey questions and drinking time from survey data. As mentioned in chapter 3.1, survey data has totally 114 survey questions, but not all of these questions are useful to my research study. Only patient ID, survey start time, survey end time, and survey questions about how many alcohol the patient drank are needed. The rows that have drinking time in the survey data also needs to be extracted. Then according to the survey start time and survey end time, the raw data can be labeled.

3.2.2 Raw Data preprocessing

1. Raw Data Labeling

The way to label the raw data is to join the raw data and the preprocessed survey together based on the time. Then label the data based on the drinking episodes. For each drinking episodes, label the raw data from 30 minutes before the episode start time to 2 hours after the end time of the drinking episode as drinking.

3.3 Data Statistics and Best User Selection

This section is going to show some basic statistics about both the survey data and the raw data. Since the number of days that each user participates in this research is different from that of each other and in some of the survey days some users do not have raw data collected, these statistics is helpful for choosing the best users.

The survey data statistics will include the number of users we have, the number of drinking days for each user, and how many drinking episodes each user has. The raw data statistics will include the number of days of raw data each user has, the total number of records of raw data each user has, the total of drinking records for each user, the ratio between the number of drinking records and the number of total records.

3.3.1 Statistics on survey Data

There are totally 29 patients participating in this project. Each user is scheduled for different number of surveys. For the survey data, each user may have drinking behavior on some survey days but not have drinking behavior on some other days. The following Figure 4 is the statistics for the number of survey days with drinking. It shows that user 2867 has the most number of survey days with drinking. And the users who has more than 10 survey days with drinking includes user 1510, 2867, 2958, 3319, 3641, 4405, 4489, 4540, 4557, and 4620.

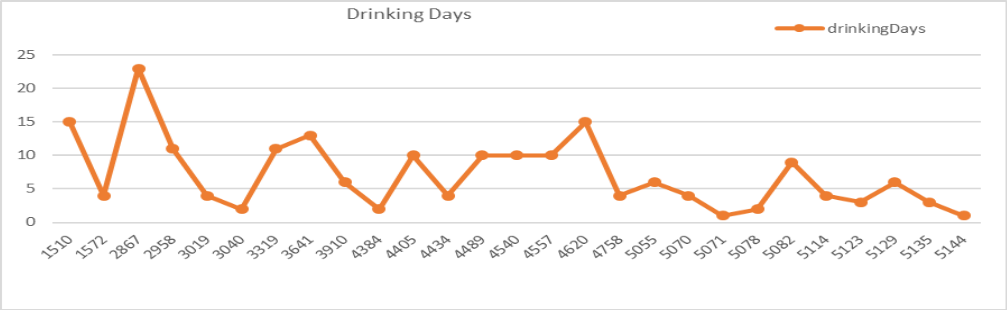


Figure 4. Survey Days with Drinking for Each User

In the surveys, each user may have multiple drinking episodes in each day. The more drinking episodes one user has in each day; the more drinking raw data the user will have. The Figure 5 is the statistic for drinking episodes for each user.

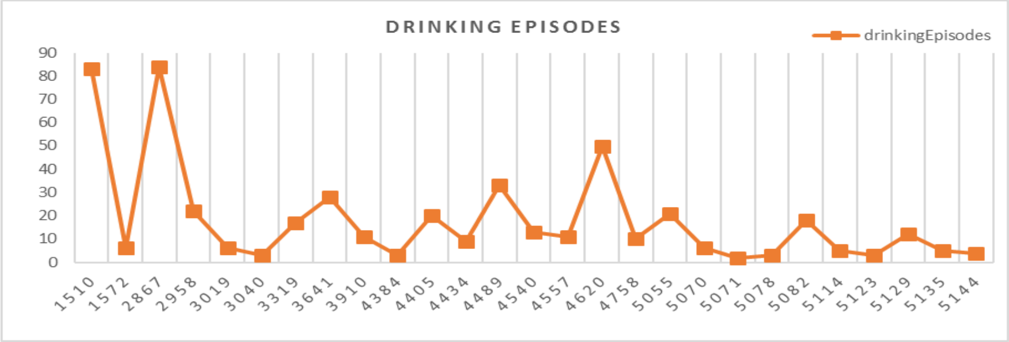


Figure 5. Drinking Episodes for Each User

User 2867 has the most number of drinking episodes. The users who have drinking episodes more than 25 includes user 1510, 2867, 3641, 4489, and 4620.

3.3.2 Statistics on Raw Data

This section will show the statistics of raw data for all the users. The raw data statistics will include the number of days that each user has raw data, the total number of raw data records each user has, the total number of raw data that are labeled with drinking each user has, and the percentage between drinking records and total raw data records.

First have a look at the number of days each user has raw data. Figure 6 is the statistics for the number of days each user has raw data.

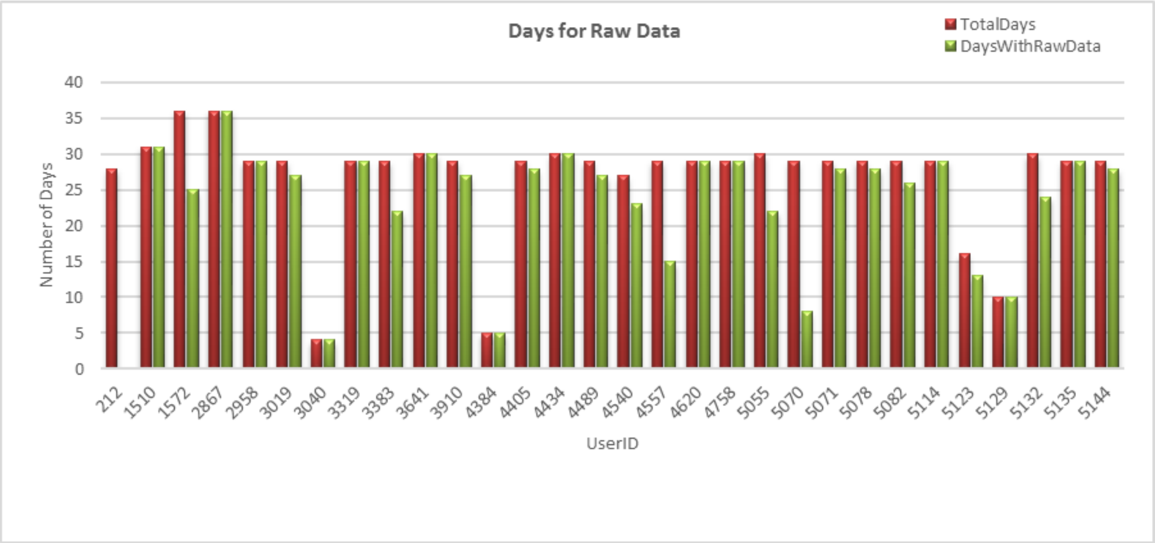


Figure 6. Number of Days Each User Has Raw Data

There are totally 29 users, but not every user has many days of raw data. For example, user 212 has 28 days of survey data, but none of these days has raw data. And user 1572 has 36 days of survey data, but it has only 25 days of raw data. Some other users have the same situations.

Secondly, take a look at the statistics of total number of raw data records for each user. As already mentioned before, the raw data is collected by every minutes. The more number of raw data records each user has, the more information that user will have. The following Figure 7 is the statistics for raw data records each user has.

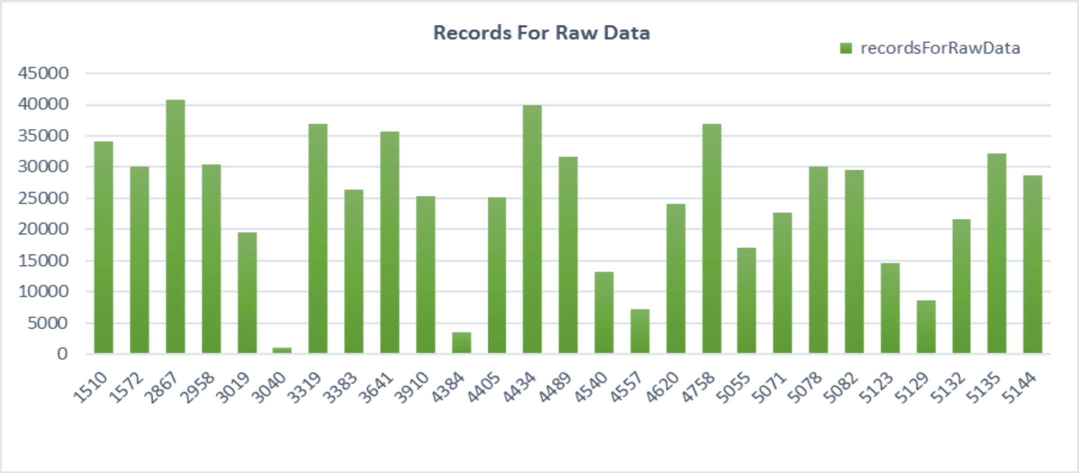


Figure 7. Number of Raw Data Records

From this figure, user 2867 and user 4434 have the most number of raw data records. And there are eight users who have more than 30,000 records of raw data. They are user 1510, 2867, 2958, 3319, 3641, 4434, 4489, 4758, and 5135. This result is consistent with the statistics for the number of days of raw data. The users who have more number of days of raw data have more number of raw data records.

Next is the statistics for the number of drinking records each user has. Drinking records means the records that are labeled with drinking. In the data preprocessing section, how to label drinking records will be introduced. The following Figure 8 is the statistics for number of drinking records each user has.

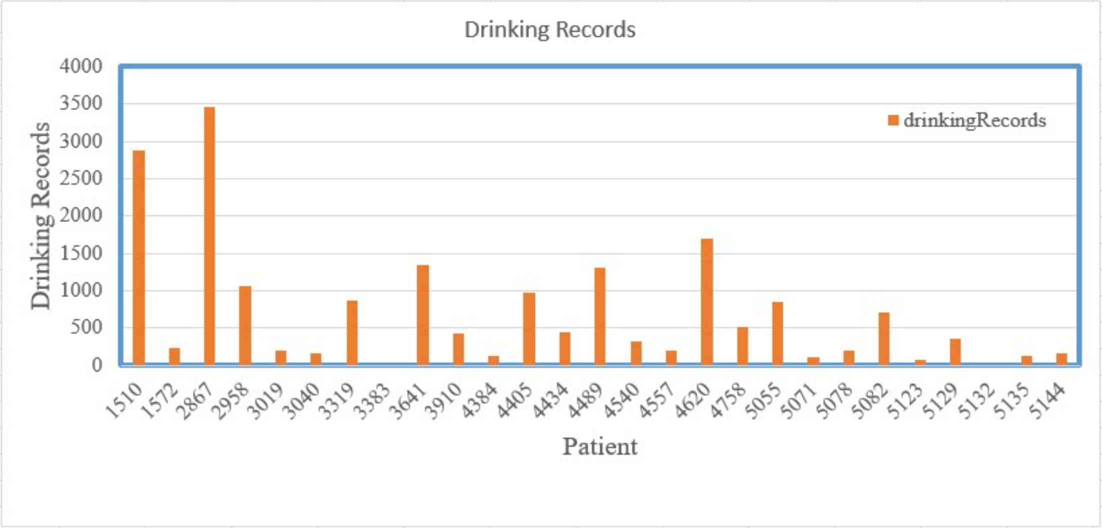


Figure 8. Number of Drinking Records

In this graph, there are five users who have more than 1000 drinking records. They are 1510, 2867, 2958, 3641, 4489, and 4620. Not all of these users have the most number of raw data records. For example, user 4620 has the top five number of drinking records, but it does not have top five number of raw data records. The number of drinking records is a very important criterion for choosing good users. Those users who have the most number of drinking records will be choose.

Now have a look at the last raw data statistics, which is the percentage between the drinking records of all users and the total number of raw data records. The total number of drinking records and the total number of raw data records are showed in the following Figure 9.

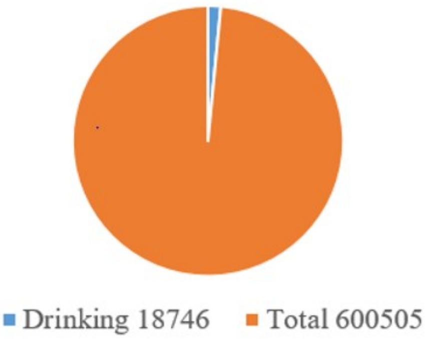


Figure 9. Number of Drinking Records and Total Raw Data Records

In figure 9, the total number of drinking records for all users is 18746, and the total number of raw data records for all users is 600505. The percentage between the total number of drinking records and the total number of raw data records is 3.12%. So although there are a large number of raw data records, the percentage of drinking records is very small.

3.3.3 Combined Statistics for Survey Data and Raw Data

This section is about the statistics for survey data and raw data. It is the matching days between survey data and raw data. Each user has many days of survey data and many days of raw data. But these two kinds of data are collected separately. So it may happen that one user has survey data on some day but he/she does not have raw data on that day. If one user has both survey data and raw data on the same day, then it is a matching day. The following Figure 10 is the statistics for matching days.

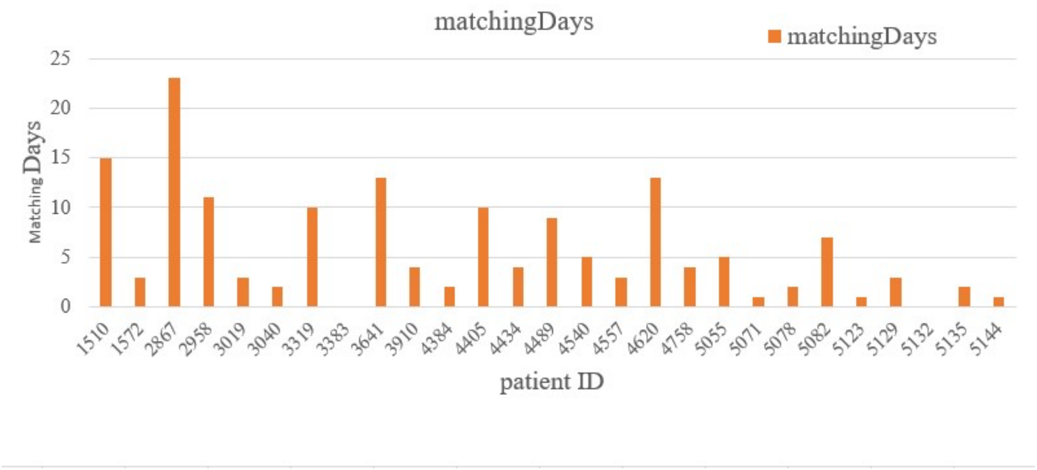


Figure 10. Number of Matching Days

There are only five users who have more than 10 matching days. But most users have more than 25 days of raw data and survey data. So each user has many of non-matching days, which means a lot of data is useless to us. If a user has more matching days, he/she may have more drinking records. All the drinking records come from these matching days.

According to the above statistics, user 2867 is the best user because he/she has the most days of data, most number of drinking records. User 2867 will used to do data cleaning. Then apply the same method to other users.

3.4 Feature Selections

This chapter is about feature selection. Feature selection will base on the feature visualization and some preliminary experimental results.

3.4.1 Data Visualization

This section is about data visualization for the features of the raw data. The data visualization is useful to discover the general pattern of the features and to deal with the raw data.

First, take a look at the visualization of skin temperature in Figure 11.

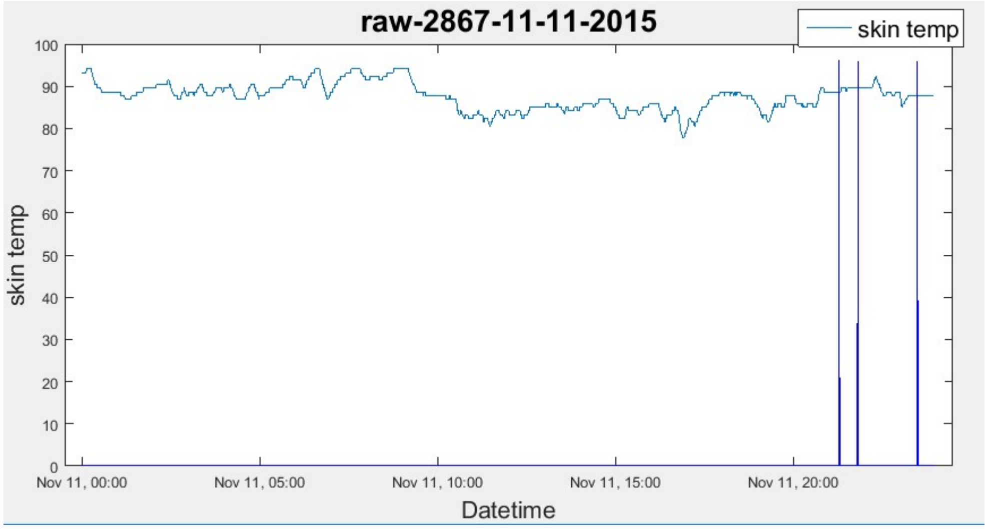


Figure 11. Visualization for Skin Temperature

The raw data in figure 11 is the skin temperature from user 2867 and this raw data was collected on November 11th, 2015. The blue vertical lines in the graph are the time when user 2867 has alcohol drinking. This figure shows that the skin temperature is fluctuating all the time. But the temperature goes up after user 2867 has alcohol drinking.

Following is the visualization of heart rate. This heart rate data is from the same day as the skin temperature data above. The blue lines in Figure 13 present the time when user 2867 drinks alcohol. And figure 12 is the normal heart rate signal.

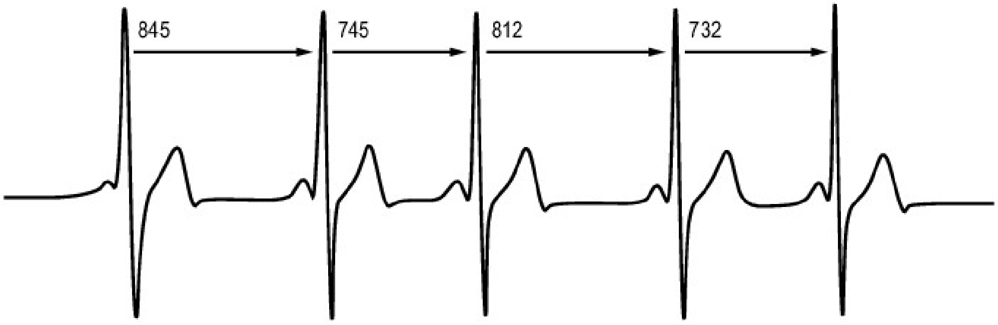


Figure 12. Normal Heart Rate Signal Chart

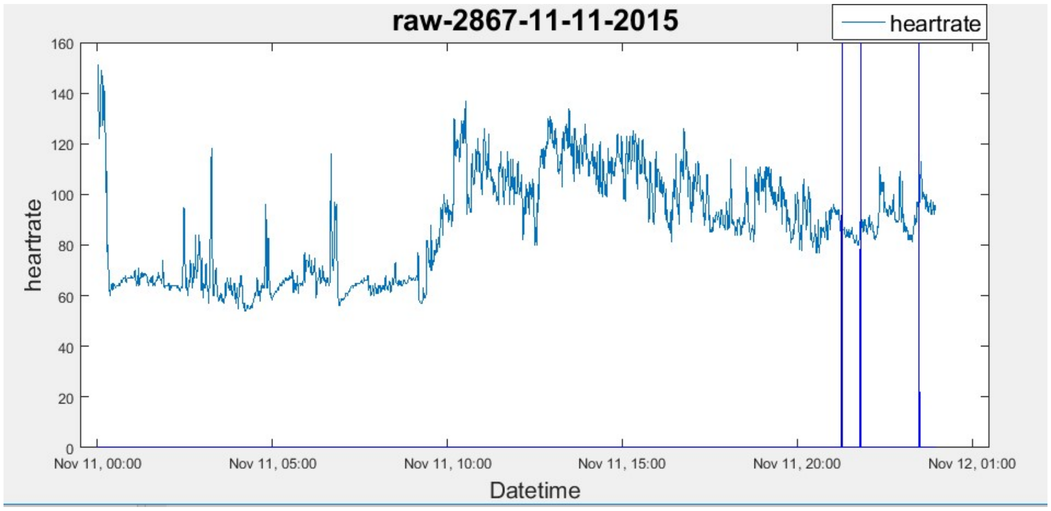


Figure 13. Visualization for Heart Rate

Comparing the user 2867's heart rate signal with the normal heart rate signal, user 2867's heart rate is very noisy. The heart rate goes up after user 2867 has some alcohol drinking, which is consistent with the physical rules.

Next will be the visualization for GSR. GSR is also known as electro dermal activity (EDA). The traditional theory of EDA holds that skin resistance varies with the state of sweat glands in the skin. Sweating is controlled by the sympathetic nervous system, and skin conductance is an indication of psychological or physiological arousal. If the sympathetic branch of the autonomic nervous system is highly aroused, then sweat gland activity also increases, which in turn increases skin conductance. In this way, skin conductance can be a measure of emotional and sympathetic responses [2]. Although there are not too many direct researches on how alcohol affects GSR, this knowledge may it helpful to understand how alcohol affects GSR.

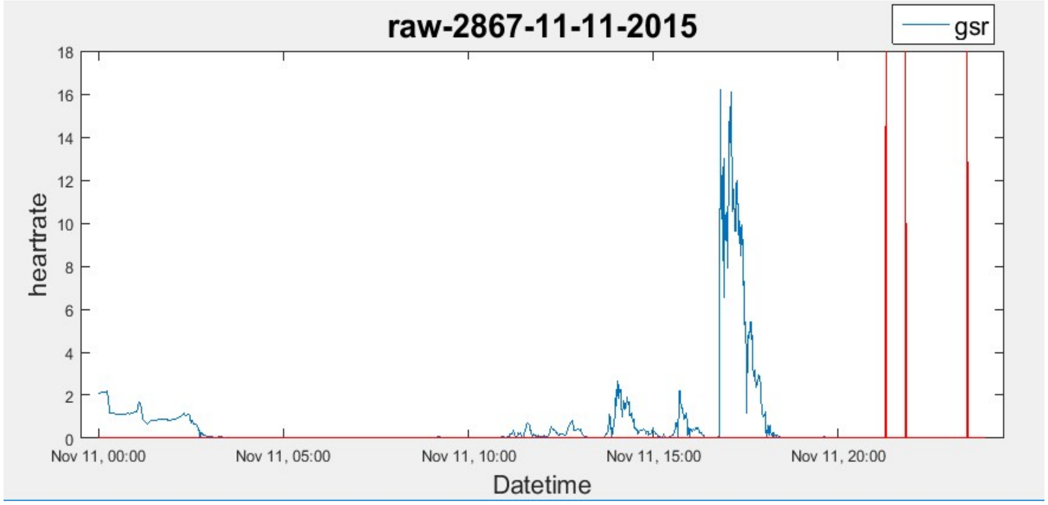


Figure 14. Visualization for GSR

In figure 14, the blue line is the signal of GSR and the red lines are the time when user 2867 has alcohol drinking. The GSR drops a lot after the drinking, but it’s not sure whether this is caused by the alcohol.

Next is the visualization for steps. Figure 15 is the visualization for steps.

The red lines in this graph are steps and the three vertical blue lines are drinking. There are many gaps in the graph, which means the patient sometimes stays and sometimes walks. During the drinking time, the patient almost does not walk, which makes sense. Because when people drink, they always sit down.

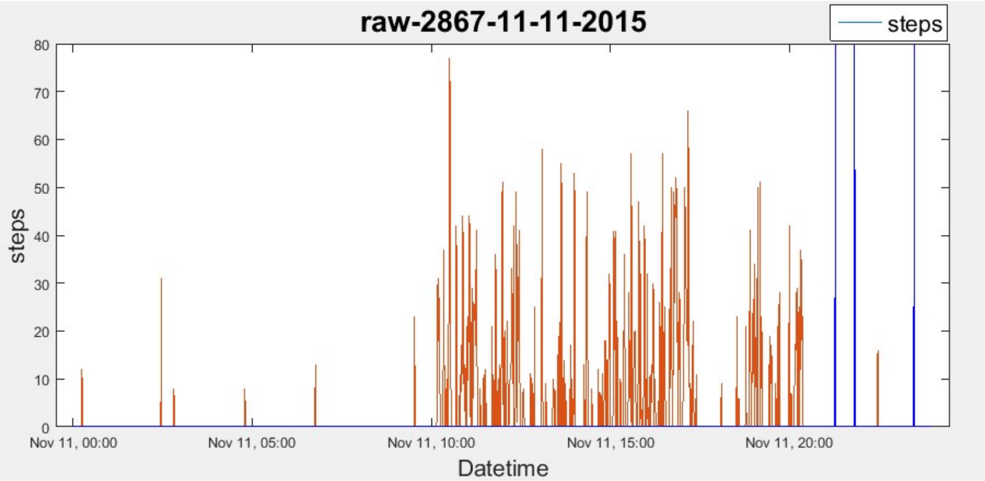


Figure 15. Visualization for Steps

Last is the visualization of calories. Figure 16 is the visualization of calories. In this graph, there are five periods and the amount of calories are the same in each period.

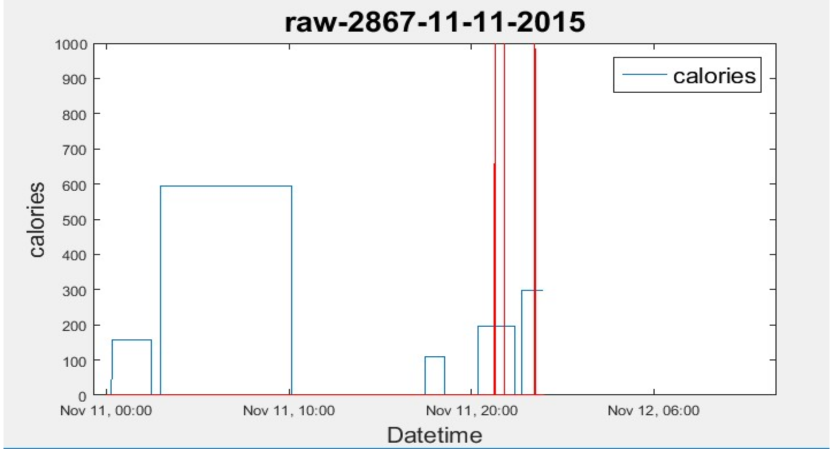


Figure 16. Visualization for Calories

The visualization of five features shows that steps and calories are very noise features. Some preliminary results will be showed to see if they are really not good features.

3.4.2 Experimental Results with and Without Steps and Calories

Table 1 is the J48 result for user 2867 with steps and calories. Table 2 is the j48 result for user 2867 without steps and calories. The accuracy in table 1 is 70.54% and the kappa value is 0.4105. The accuracy in table 2 is 82.93% and the kappa value is 0.6586. So the result without steps and calories is much better than the result with steps and calories, which further illustrates that steps and calories should be discarded. So three features, heart rate, skin temperature and GSR will be used in later work.

Table 1. Confusion Matrix for 2867 with Steps and Calories

		Predicted	
		0	1
Actual	0	962	228
	1	474	685

Table 2 Confusion Matrix for 2867 without Steps and Calories

		Predicted	
		0	1
Actual	0	976	214
	1	187	972

3.5 Data Cleaning

This section is going to talk about the data cleaning. The data cleaning process will include filling missing value, removing extreme value and outliers, smoothing, and removing gaps.

1 Remove Insufficient Data and Gaps

Some data has many missing data points. Within a 10-minuteIf size moving window, if half of the data is missing, then it will be considered as insufficient data. Figure 17 is an example of insufficient data. If there is no data in 10-minute size moving window, then it is a gap and it will be removed. Figure 18 is an example of gaps.

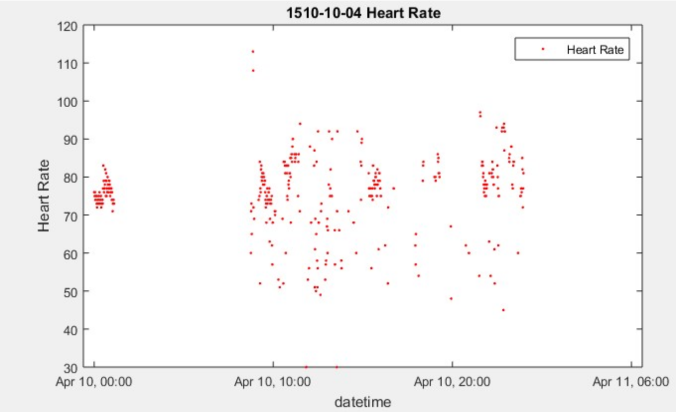


Figure 17. Insufficient Data Example

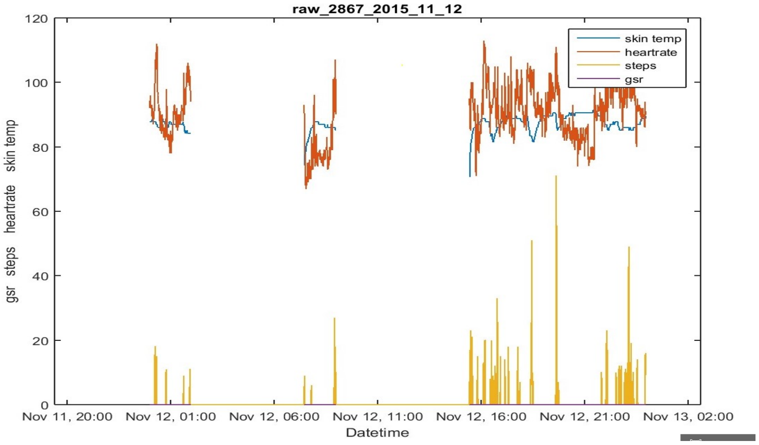


Figure 18. Gap Example for User 2867 11/11/2015

2 Removing Outliers

The raw data is very noisy and there are a lot of outliers. Outliers have a very big effect on the feature extraction, especially on mean. So they need to be removed. Loess can be used here. 1 percent of the data is used as the span parameter of loess. In this span any data points that is smaller than two standard deviations below the mean or greater than two standard deviations above the mean, will be considered as outliers. Figure 19 is an example of removing heart rate outliers.

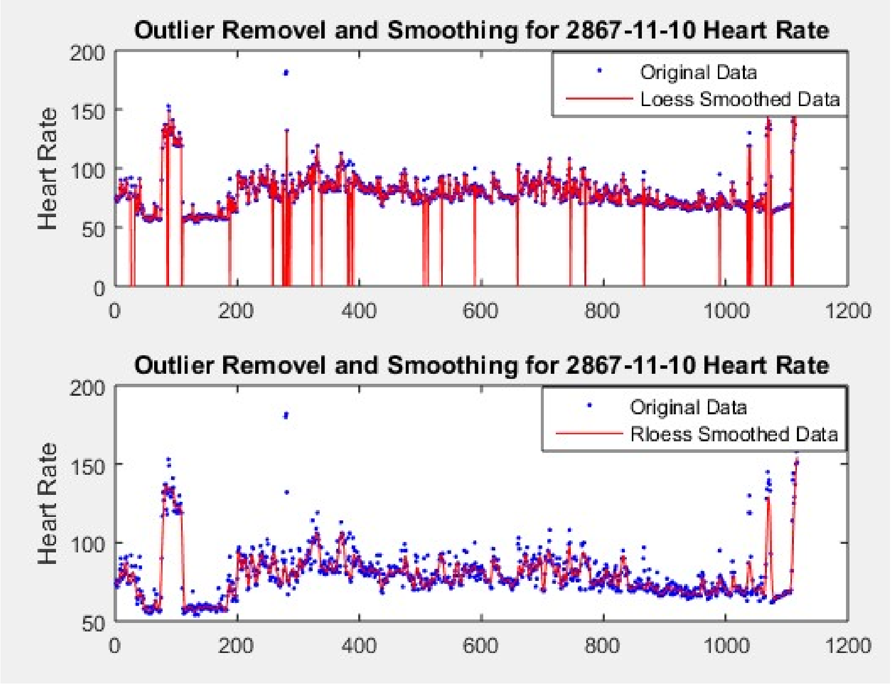


Figure 19. Outliers Removal Example

In Figure 19 there are two outliers removing methods. The above one is using loess and the below one is using robust loess. The robust loess has a better result in removing outliers.

3 Data Smoothing

Robust loess is used to smooth the data. The robust loess uses locally weighted linear regression to smooth the data. And it can detect the outliers and does not use the outliers to smooth the data. Figure 20 is the plot of heart rate before smoothing and Figure 21 is the corresponding plot after robust loess smoothing. The smoothed data is much more smoothing than the raw data. The robust smoothing procedure follows the following steps:

1) Calculate the residuals for each points in the span
2) Compute the robust weights for each data in the span, the weights are given by the following function:

$$w_i = \begin{cases} \left(1 - (r_i/6MAD)^2\right)^2, & |ri| < 6MAD, \\ 0, & |ri| \geq 6MAD, \end{cases}$$

Where ri is the resudual of the ith data point produced by the regression, and MAD is the median absolute deviation of the residuals.

$$MAD = \text{median}(|r|).$$

3) Smooth the data using the robust weights

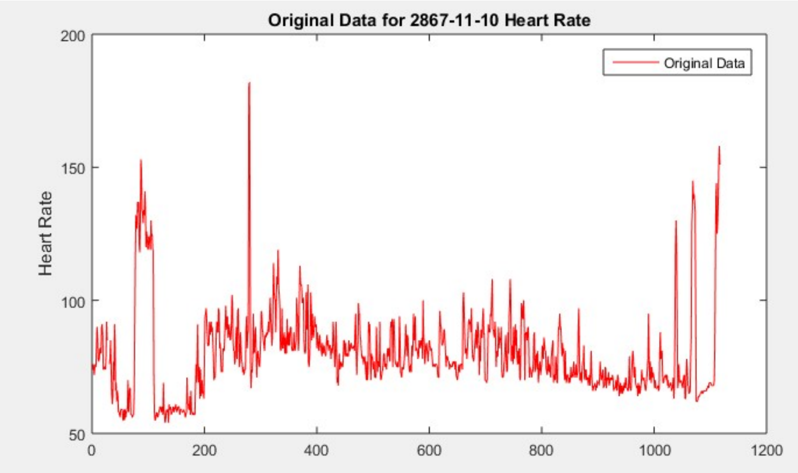


Figure 20. Heart Rate Plot before Loess

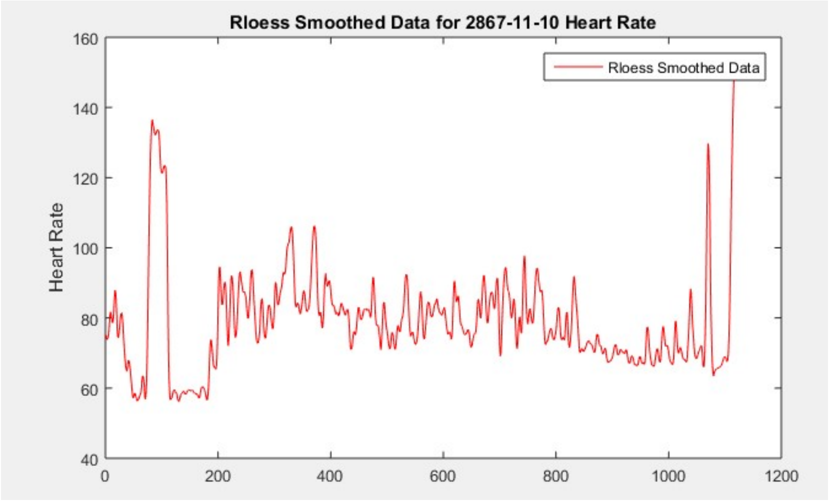


Figure 21. Heart Rate Plot after Loess

4. Data Analysis Pipelines

This chapter is going to introduce the implementations of three pipelines. The first pipeline is drinking record prediction pipeline. The second one is statistical drinking episode prediction pipeline. This pipeline will extract statistical features from one dimensional raw feature first, then use machine learning methods, like Naïve Bayes, Beyes Network, Logistic Regression, and J48 decision tree, to do the classification on the statistical Features. The third one is to use deep learning, cifar 10, to extract deep learning features from the raw data, and then apply machine learning methods on the deep learning features.

4.1 Drinking Records Prediction Pipeline

Figure 22 is the whole process for the drinking records prediction pipeline.

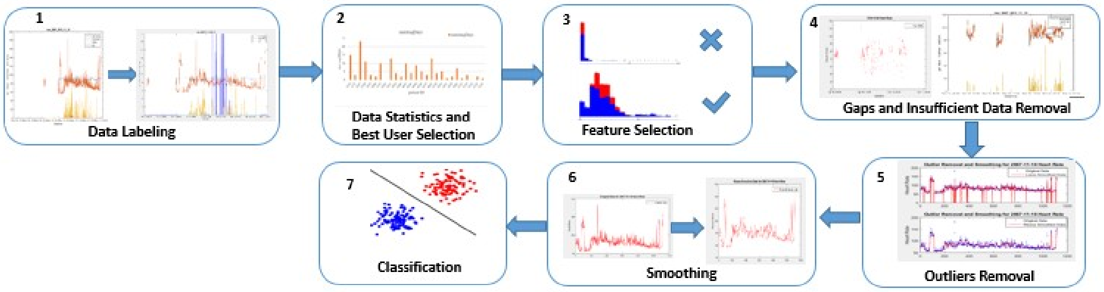


Figure 22. Drinking Records Prediction Pipeline

This pipeline includes data preprocessing, data statistics and best user selection, feature selection, data cleaning, and classification. The first four steps, data preprocessing, data statistics and best user selection, feature selection, data cleaning have already been introduced in chapter 3. So it will not be repeat here. The classification step uses four classifiers, Naïve Bayes, Beyes Network, Logistic Regression, and J48 decision tree. Following are the results for drinking records prediction pipeline on both raw data and cleaned data.

1. Result for cleaned data

Table 3 is the result for drinking record prediction with cleaned data. The best classifier is Bayes Network. The accuracy is 71.87% and the Kappa value is 0.4373.

Table 3. Confusion Matrix for Drinking Record Prediction Cleaned Data

		Predicted	
		0	1
Actual	0	4229	822
	1	2020	3031

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	63.47%	71.87%	69.09%	65.25%

2. Result for Raw Data

Table 4 is the result for drinking record prediction with raw data. The best classifier is J48. The accuracy is 80.67% and the Kappa value is 0.6131.

Table 4. Confusion Matrix for Drinking Record Prediction Raw Data

		Predicted	
		0	1
Actual	0	4232	870
	1	1083	3917

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	69.42%	78.09%	70.07%	80.67%

3. Result Comparison between Cleaned Data and Raw Data

Comparing the result for cleaned data with the result for raw data, the result for raw data is much better than the cleaned data. So the raw data other than cleaned data will be used for other two pipelines.

4.2 Drinking Episodes Prediction Statistical Pipeline

In this drinking episodes prediction statistical pipeline, statistical features, like mean, standard deviation, skewness, slope, and coefficient of variation, are extracted from one dimensional signal. Then apply machine learning classifier on these statistical features. The following figure is the whole steps of this pipeline. The details for each step will be explained.

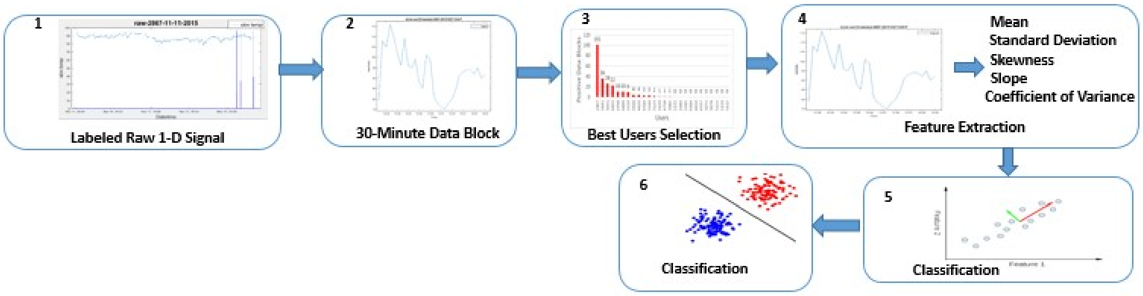


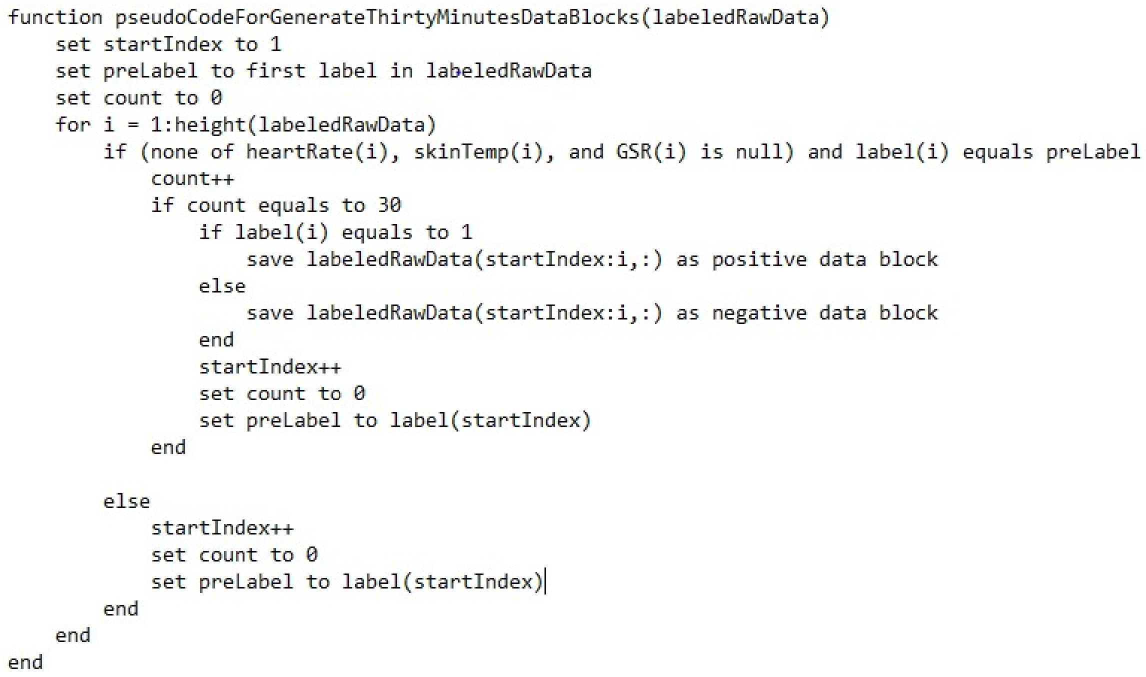
Figure 23. Drinking Episodes Prediction Statistical Pipeline

1. Raw Single 1-D Signal

The three features, heart rate, skin temperature, and GSR, are all one dimensional signal. So we call them 1-D signal. I am going to use these three 1-D signal to generate statistical features in step four.

2. Generate 30 Minutes Data Block

As already introduced in chapter 4.1, drinking prediction based on each one-minute record is not accurate because there is overlapping information between near records. In order to eliminate the overlapping information, the original raw data is split into 30 minutes data blocks. Then there won't be any overlapping information between any two data blocks. Figure 24 is the pseudo code for generating 30-minute data blocks.



```
function pseudoCodeForGenerateThirtyMinutesDataBlocks(labeledRawData)
    set startIndex to 1
    set preLabel to first label in labeledRawData
    set count to 0
    for i = 1:height(labeledRawData)
        if (none of heartRate(i), skinTemp(i), and GSR(i) is null) and label(i) equals preLabel
            count++
            if count equals to 30
                if label(i) equals to 1
                    save labeledRawData(startIndex:i,:) as positive data block
                else
                    save labeledRawData(startIndex:i,:) as negative data block
                end
                startIndex++
                set count to 0
                set preLabel to label(startIndex)
            end

        else
            startIndex++
            set count to 0
            set preLabel to label(startIndex)
        end
    end
end
```

Figure 24. Pseudo Code for Generating 30-Minute Data Blocks

First according to the survey question, find all the drinking time for each user. If the time difference between two drinking behaviors is shorter than two hours, it is considered as one drinking episode. Each drinking episode has a start time and end time. All the data points between half hour before the drinking episode start time and two hours after the drinking episode end time will be considered as drinking records. Other data points will be considered non-drinking. After labeling the raw data, divide the raw data into two kinds of data blocks. One is positive and the other is negative. Scan through the raw data from the first record until 30 records. If the 30 minutes data block contains all drinking records, then it is positive. If the 30 minutes data block contains all non-drinking records, then it is negative. Each data block contains only drinking records or non-drinking records. If there is missing value, then begin from the data point right after the missing value. Repeat above steps until we finish all the all data.

Figure 25 is the visualization of positive data block.

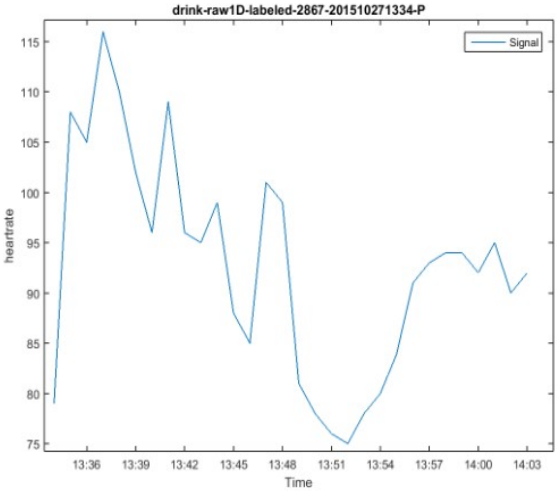


Figure 25. Visualization of Positive Data Block

Figure 26 is the visualization of negative data block.

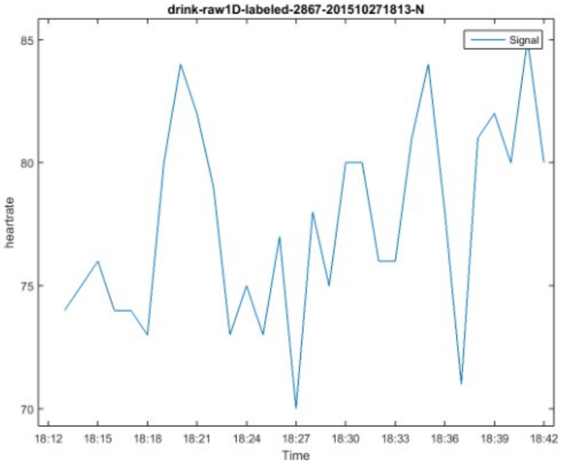


Figure 26. Visualization of Negative Data Block

3. Best Users Selection

Although there are 29 users in this study, many users have very few drinking days. When generating the thirty-minute data blocks, many people have even fewer positive data blocks. And the result of classification on small number of data will not be accurate. So it is important to choose the users who have the most number of positive data blocks. The following figure is the statistics for the number of positive data blocks for all users. Users 2867, 3641, and 5055 have the most positive data blocks. They will be used in later experiment.

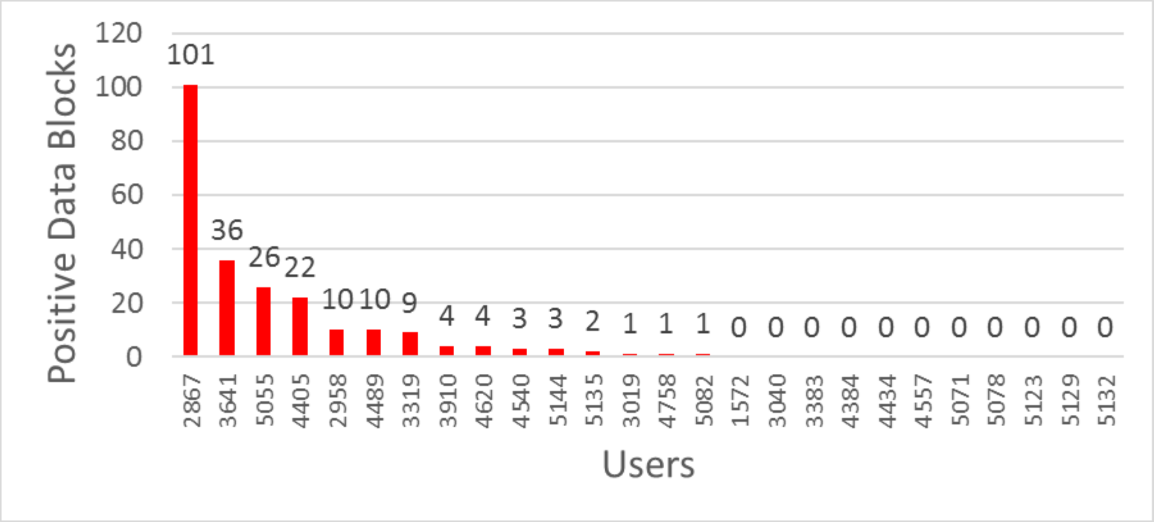


Figure 27. Number of Positive Data Blocks

4. Statistical Features Extraction

After generating the 30 minutes data blocks, it is time to extract statistical features from these data blocks. The following statistical features will be extracted.

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 the linear regression fitted on the thirty-minute data block

Coefficient of Variance: Std/Mean

5. Classification 1

After extracting the statistical features, then apply the following machine learning classifiers to them.

- Naïve Bayes
- Bayes Network
- Logistic Regression
- J48 Decision Tree

6. Principal Component Analysis

Totally 15 features are extracted from the three raw features. Some of these extracted features may have correlation between each other. So apply principal component analysis on them to eliminate the correlation.

7. Classification 2

After applying principal component analysis on the extracted features, apply the same classifiers described in step 6 again. Then compare the result before principal component analysis and after principal component analysis.

4.3 Drinking Episodes Prediction Deep Learning Pipeline

This chapter is going to introduce the drinking episode prediction deep learning pipeline. This pipeline will generate spectrogram from the 30 minutes data blocks. Then use deep learning model, Cifar 10, to extract features from the spectrogram. Finally apply machine learning classifier on the deep learning features. Figure 28 is the whole process of this pipeline.

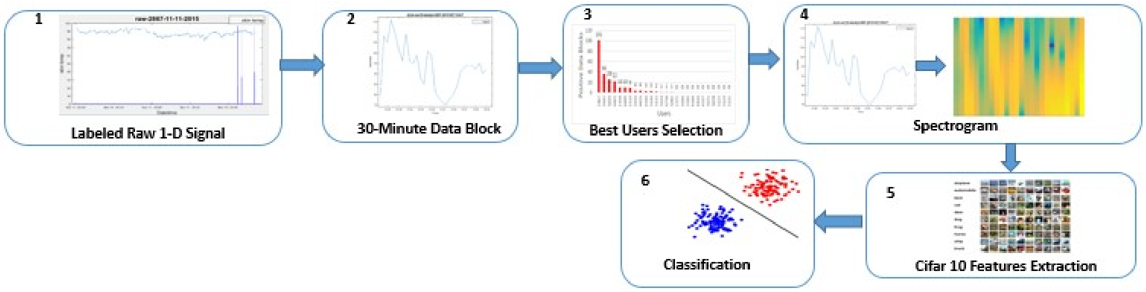


Figure 28. Drinking Episodes Prediction Deep Learning Pipeline

This pipeline has 6 steps. The first three steps are the same as the drinking episode prediction statistical pipeline. So jump to step 4 directly.

Step 4 is to generate spectrogram from the 30 minutes data blocks. Each data block will have a corresponding spectrogram. Because deep learning is very good at extract information from graphs, we convert the 1-D signal into spectrogram. Figure 29 is an example of spectrogram generated from user 2867.

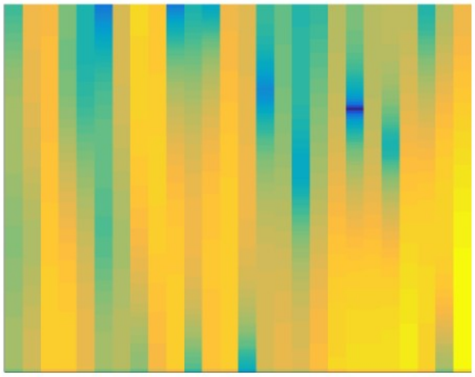


Figure 29. Spectrogram drink_normalize_d_2DH_2867_201511101351_P

Step 5 is to use deep learning model Cifar 10 to extract deep learning features from the spectrogram. Cifar 10 is an existing deep learning model. It has 10 categories of animals and uses these animals to simulate the image and get a weight for each of the 10 animals. Then use these weights to do classification.

The last step is to apply machine learning classifiers on the deep learning features extracted from step 5. The machine learning classifiers used in this pipeline are the same as those classifiers in the drinking episodes prediction statistical pipelines.

5. Experimental Design and Results

After introducing the implementation of all pipelines, this chapter will talk about the experimental design and experimental results.

5.1 Experimental Design

1. Training Data and Testing Data Design

Based on the data statistics from chapter 2, the percentage of the drinking records between the total records is 3.12, which is very small. When doing the drinking prediction, if all the instances are classified as non-drinking, then the accuracy will be 96.88%. So in this case the baseline will be 96.88%, which is too high. Even if classifying all the instances correctly with 100% accuracy, it is only a little bit higher than the baseline. This is not the result we want. In order to solve this problem, this paper use balanced training data set and testing data set. Balanced means the number of positive instances and the number of negative instances in both training data set and testing data set are the same. For example, in the statistical pipeline for drinking episode prediction, users 2867 has 101 positive thirty-minute data blocks in total. Then randomly select 101 negative thirty-minute data blocks from the negative thirty-minute data blocks pool. Then the 101 positive thirty-minute data blocks and the 101 negative thirty-minute data blocks will be the training and testing data set. If 66% of this data set is used as training and the rest is used as testing, then 67 instances from the positive data blocks pool and 67 instances from the negative data blocks pool will form the training data set. And 34 instances from the positive data blocks pool and 34 instances from the negative data blocks pool will be the testing data set. In this case, the baseline accuracy will be 50%, which is more reasonable.

2. Within-User Experimental Design

Different users may have different physiological features from each other. And their bodies may have different reactions to alcohol. So the experimental data is very diverse between different users. It may not be able to find a model that can apply to all users. So in order to eliminate the difference between different users, this work will try to find different models for different users. In other words, the training data set and the testing data set are from the same users. For example, user 2867 has 101 positive and 101 negative thirty-minute data blocks. In order to find the best classifier only 2867, use 66% of these 202 instances as training and 34% of them as testing. The same rule will apply to other users.

3. Between-Users Experimental Design

Between-users experimental design aims to find a general model that applies to all users. So the training data set and the testing data set are from different users. For example, assume the goal is to find a model that applies to user 2867 and user 2958. User 2867 has 101 positive and 101 negative thirty-minute data blocks. User 2958 has 39 positive and 39 negative thirty-minute data blocks. Then combine the 202 data instances from user 2867 and the 78 data instances from user 2958. 66% of these combined data will be used as training data set and the rest 34% will be used as testing. This is how the between-users experiments are designed.

4. Performance Measurement

The performance measurement will be a judgement for whether the result is good or not. The performance measurements that are used in this work are: confusion matrix, accuracy, and kappa. Assume that a, b, c, d represent true positive, false positive, false positive, and true negative. The following are the formulas to calculate the kappa statistic.

$$p_o = \frac{a+d}{a+b+c+d}$$

$$p_{\text{Yes}} = \frac{a+b}{a+b+c+d} \cdot \frac{a+c}{a+b+c+d}$$

$$p_{\text{No}} = \frac{c+d}{a+b+c+d} \cdot \frac{b+d}{a+b+c+d}$$

$$p_e = p_{\text{Yes}} + p_{\text{No}}$$

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Pe is the probability of random agreement, and Po is the observed proportionate agreement.

5.2 Experimental Results

This chapter is about the experimental results for the statistical pipeline and the deep learning pipeline. For the statistical pipeline, there will be results for skin temperature only, heart rate only, GSR only and all three features together. Also there is experimental results for statistical pipeline with principle component analysis. For the deep learning pipeline, there will be skin temperature only, heart rate only and GSR only. And both within-user results and cross-users results will be included. The within-user results will include results from the best three users 2867, 3641, and 5055 separately. The cross-users results will include results from all the data from these three users. For each individual experimental case, four classifiers are trained. They are Naïve Bayes, Bayes Network, Logistic, and J48 decision tree. Since there are too many experimental results, only one confusion matrix for the best classifier is going to be shown, but the accuracies for other classifiers will be included.

5.2.1 Statistical Pipeline

5.2.1.1 Skin Temperature Features Only for Statistical Pipeline

1. Skin Temperature Within-User 2867

The following table is the experimental result for user 2867 using only skin temperature features. The result shows that Bayes Network is the best classifier among the four classifiers. The accuracy is 63.77% and the Kappa value is 0.2191. The true positive value is 0.947, which is very good. And the true negative value is 0.258.

Table 5. Confusion Matrix Skin Temperature Within-User 2867

		Predicted	
		0	1
Actual	0	8	23
	1	2	36

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	43.49%	63.77%	47.83%	44.93%

2. Skin Temperature Within-User 3641

The following table is the experimental result for user 3641 using only skin temperature. From the comparison between the four classifiers, Logistic and Naive Bayes have the highest accuracy of 58.33%. Logistic has a Kappa value of 0.0625, but Naive Bayes has a smaller kappa value of 0.0323. So Logistic is the best classifier for user 3641 with skin temperature. The true positive value for Logistic is 0.857. And the true negative value for Logistic is 0.200.

Table 6. Confusion Matrix Skin Temperature Within-User 3641

		Predicted	
		0	1
Actual	0	2	8
	1	2	12

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	58.33%(0.0323)	50%	58.33%	54.17%

3. Skin Temperature Within-User 5055

The following table is the experimental result for user 5055 using only skin temperature features. Both Naive Bayes and J48 have the highest accuracy of 88.89%. But take a look at the Kappa value, Naive Bayes has a Kappa value of 0.7692, which is bigger than the Kappa value 0.75 of J48. So Naive Bayes is the best classifier among the four classifiers. The true positive value for Naive Bayes is 0.833. And the true negative value for Naive Bayes is 1.00.

Table 7. Confusion Matrix Skin Temperature Within-User 5055

		Predicted	
		0	1
Actual	0	6	0
	1	2	10

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	88.89%	55.56%	61.11%	88.89%(0.75)

4. Skin Temperature Cross-User

The following table is the experimental result for cross users 2867, 3641, and 5055 using only skin temperature features. Naive Bayes is the best classifier among the four classifiers with an accuracy of 63.06%. The Kappa value is 0.2501. The true positive value is 0.895. And the true negative value is 0.352.

Table 8. Confusion Matrix Skin Temperature Cross-User

		Predicted	
		0	1
Actual	0	19	35
	1	6	51

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	63.06%	48.65%	61.26%	54.96%

5.2.1.2 Heart Rate Features Only for Statistical Pipeline

1. Heart Rate Within-User 2867

The following table is the experimental result for user 2867 using only heart rate features. Naive Bayes is the best classifier among the four classifiers, which has the highest accuracy of 81.16%. The Kappa value for Naive Bayes is 0.6112. The true positive value for it is 0.921, which is very good. And the true negative value is 0.677.

Table 9. Confusion Matrix Heart Rate Within-User 2867

		Predicted	
		0	1
Actual	0	21	10
	1	3	35

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	81.16%	72.46%	76.81%	76.81%

2. Heart Rate Within-User 3641

The following table is the experimental result for user 3641 using only heart rate features. Logistic is the best classifier among the four classifiers, because it has the highest accuracy of 66.67%. The Kappa value for it is 0.2941. The true positive value for it is 0.786 and the true negative value is 0.5.

Table 10. Confusion Matrix Heart Rate Within-User 3641

		Predicted	
		0	1
Actual	0	5	5
	1	3	11

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	58.33%	62.5%	66.67%	58.33%

3. Heart Rate Within-User 5055

The following table is the experimental result for user 5055 using only heart rate features. Compared to other three classifiers, J48 has the highest accuracy of 88.89%. So J48 is the best model among the four models. The Kappa value for it is 0.75. The true positive value for it is 0.917 and the true negative value is 0.833.

Table 11. Confusion Matrix Heart Rate Within-User 5055

		Predicted	
		0	1
Actual	0	5	1
	1	1	11

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	83.33%	55.56%	83.33%	88.89%

4. Heart Rate Cross-User

The following table is the experimental result for cross-users 2867, 3641, and 5055 using only heart rate features. Naive Bayes has the highest accuracy of 75.68%. The Kappa value for Naive Bayes is 0.5096. The true positive value for it is 0.895 and the true negative value is 0.611.

Table 12. Confusion Matrix Heart Rate Cross-User

		Predicted	
		0	1
Actual	0	33	21
	1	6	51

5.2.1.3 GSR Features Only for Statistical Pipeline

1. GSR Within-User 2867

The following table is the experimental result for user 2867 using only GSR features. The experimental results shows that J48 is the best classifier, which has the highest accuracy of 75.36%. The Kappa value for J48 is 0.5094. The true positive value for it is 0.711. And the true negative value is 0.806.

Table 13. Confusion Matrix GSR within-User 2867

		Predicted	
		0	1
Actual	0	25	6
	1	11	27

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	63.77%	72.46%	65.22%	75.36%

2. GSR Within-User 3641

The following table is the experimental result for user 3641 using only GSR features. Both Naive Bayes and Logistic have the highest accuracy, which is 70.83%. Then take a look at the Kappa value, Naive Bayes has a Kappa value of 0.4085. However, Logistic has a bigger Kappa value 0.44. So Logistic is the best classifier among the four classifiers. The true positive value for Logistic is 0.571. And the true negative value for Logistic is 0.900.

Table 14. Confusion Matrix GSR within-User 3641

		Predicted	
		0	1
Actual	0	9	1
	1	6	8

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	70.83%(0.4085)	41.67%	70.83%	45.83%

3. GSR Within-User 5055

The following table is the experimental result for user 5055 using only GSR features. Naïve Bayes is the best classifier, which has the highest accuracy of 83.33%. And other three classifiers have a much lower accuracy than Naïve Bayes. The Kappa value for Naïve Bayes is 0.5714. The true positive value for it is 1.00. And the true negative value is 0.5.

Table 15. Confusion Matrix GSR within-User 5055

		Predicted	
		0	1
Actual	0	3	3
	1	0	12

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	83.33%	33.33%	38.89%	38.89%

4. GSR Cross-User

The following table is the experimental result for cross-users 2867, 3641, and 5055 using only GSR features. J48 has the highest accuracy of 60.36%. The accuracy difference between J48 and other classifiers is not too big. The Kappa value for J48 is 0.1925. The true positive value for it is 0.930and the true negative value is 0.259.

Table 16. Confusion Matrix GSR Cross-User

		Predicted	
		0	1
Actual	0	14	40
	1	4	53

5.2.1.4 All Features for Statistical Pipeline

This chapter is about the experimental results for using all the 15 extracted features from skin temperature, heart rate, and GSR features. And the experimental results will also include both within-user and cross-users. One thing that needs to be mentioned here is that SHG will be used to represent skin temperature, heart rate, and GSR in and after this chapter.

1. SHG Within-User 2867

The following table is the experimental result for user 2867 using SHG features. From the experimental results for four classifiers, J48 is the best classifier, which has the highest accuracy of 79.71%. The Kappa value for J48 is 0.5851. The true positive value for it is 0.868. And the true negative value is 0.710.

Table 17. Confusion Matrix SHG Within-User 2867

		Predicted	
		0	1
Actual	0	22	9
	1	5	33

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	72.46%	73.91%	76.81%	79.71%

2. SHG Within-User 3641

The following table is the experimental result for user 3641 using SHG features. Naïve Bayes and J48 have the same and highest accuracy 58.33% in the four classifiers. But J48 has a Kappa value of 0.1176 and Naïve Bayes has a smaller Kappa value 0.0909. So J48 is the best classifier here. And J48 has a true positive value of 0.714 and a true negative value of 0.400.

Table 18. Confusion Matrix SHG Within-User 3641

		Predicted	
		0	1
Actual	0	4	6
	1	4	10

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	58.33%(0.0909)	54.17%	54.17%	58.33%

3. SHG Within-User 5055

The following table is the experimental result for user 5055 using SHG features. Naïve Bayes and J48 have the same accuracy of 88.89%, which is the highest in the four classifiers. And 88.89% is much higher than the accuracy of other two classifiers. Since Naïve Bayes has a Kappa value of 0.7692, which is higher than the Kappa value of J48

0.75, it is the best classifier for user 5055 on SHG features. And the true positive value for Naïve Bayes is 0.833 and the true negative value for it is 1.00.

Table 19. Confusion Matrix SHG Within-User 5055

		Predicted	
		0	1
Actual	0	6	0
	1	2	10

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	88.89%	55.56%	61.11%	88.89%(0.75)

4. SHG Cross-User

The following table is the experimental result for cross-users 2867, 3641, and 5055 using only SHG features. J48 has the highest accuracy of 75.68%. The overall accuracy difference between the four classifiers is not too big. The Kappa value for J48 is 0.5158. The true positive value for it is 0.667 and the true negative value is 0.852.

Table 20. Confusion Matrix SHG Cross-User

		Predicted	
		0	1
Actual	0	46	8
	1	19	38

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	67.57%	72.97%	72.07%	75.68%

5.2.2 Statistical Pipeline with PCA

Since there are 15 extracted features, there may be correlation between some of them. Moreover, the size of the data is relatively small. So principle component analysis will be used to reduce the dimension of the features. This chapter will include the experimental results after principle component analysis.

1. PCA on SHG Within-User 2867

The following table is the experimental result for user 2867 using SHG features after PCA. Naïve Bayes is the best classifier, which has the highest accuracy of 75%. The Kappa value for it is 0.5151. The true positive value for it is 0.920. And the true negative value is 0.651.

Table 21. Confusion Matrix after PCA on SHG Within-User 2867

		Predicted	
		0	1
Actual	0	28	15
	1	2	23

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	75%	69.12%	70.59%	66.18%

2. PCA on SHG Within-User 3641

The following table is the experimental result for user 3641 using SHG features after PCA. Both Bayes Network and Logistic have the highest accuracy of 66.67%. By comparing their Kappa value, Logistic is the best classifier because its Kappa value 0.3333 is higher than that of Bayes Network, which is 0.25. The true positive value for Logistic is 0.625. And the true negative value is 0.75.

Table 22. Confusion Matrix after PCA on SHG Within-User 3641

		Predicted	
		0	1
Actual	0	6	2
	1	6	10

3. PCA on SHG Within-User 5055

The following table is the experimental result for user 5055 using SHG features after PCA. Both Bayes Network and J48 have the highest accuracy, which is 88.2353%, in the four classifiers. And they have same confusion matrix and Kappa value. So either one of them can be used as the best classifier of user 5055 on SHG features with PCA. The Kappa value for Bayes Network and j48 is 0.7639. The true positive value for them is 0.889. And the true negative value is 0.875.

Table 23. Confusion Matrix after PCA on SHG Within-User 5055

		Predicted	
		0	1
Actual	0	7	1
	1	1	8

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	82.35%	88.24%	70.59%	88.24%

4. PCA on SHG Cross-Users

The following table is the experimental result for cross-users 2867, 3641, and 5055 using SHG features after PCA. J48 is the best classifier because it has the highest accuracy of 68.18% among the four classifiers. The Kappa value for J48 is 0.3636. The true positive value for it is 0.818 and the true negative value is 0.545.

Table 24. Confusion Matrix after PCA on SHG Cross-User

		Predicted	
		0	1
Actual	0	30	25
	1	10	45

Classifier	Naïve Bayes	Bayes Net	Logistic	J48
Accuracy	65.45%	61.82%	57.27%	68.18%

5.2.3 Deep Learning Pipeline

Table 25. Deep Learning Result

Features Users	Skin Temp Features	Heart Rate Features	GSR Features
2867	44.93% J48	55.07% Logistic	63.77% Logistic
3641	50% J48	54.16% Naïve Bayes	70.83% Logistic
5055	50%	55.56%	50%

	Naïve Bayes	Logistic	J48
Cross-Users	58.59% Naïve Bayes	55.86% Logistic	63.96% Logistic
Within-User Average	48.31%	54.93%	61.53%

Table 25 is the result for deep learning. For deep learning pipeline, the overall result is not good because the best result is around 70% accuracy. The result for cross-users is better than the result for with-in user. The GSR features have better result than heart rate and skin temperature features.

5.3 Experimental Result Comparison and Analysis

This chapter will include comparison between different experimental results. The comparisons will include: comparison between features, comparison between users, and comparison between different pipelines.

1. Comparison between Features

From the following table, the results show that except user 3641, heart rate has a higher accuracy than other two features. And heart rate has the best average result for both within-user and cross-users. So heart rate is more useful than other features in drinking episode prediction. And another thing is that even for the same user the best model is different on different features.

Table 26. Statistical Pipeline Experimental Results Comparison between Users and Features

Features Users	Skin Temp Features	Heart Rate Features	GSR Features	SHG Features
2867	63.77% Bayes Net	81.16% Naïve Bayes	75.36% J48	79.71% J48
3641	58.33% Logistic	66.67% Logistic	70.83% Logistic	58.33% J48
5055	88.89% Naïve Bayes	88.89% J48	83.33% Naïve Bayes	88.89% Naïve Bayes
Cross-Users	63.06% Naïve Bayes	75.68% Naïve Bayes	60.36% J48	75.68% J48
Within-User Average	70.33%	79.14%	76.51%	75.64%
Overall Average	68.51%	78.1%	72.47%	75.65%

2. Comparison between Users

From the above table, three conclusions can be drawn by comparing the experimental results between each user. One is that user 5055 has the best result than other two users. The second conclusion is that the within-user result is better than cross-users result. The third conclusion is that different users have different best models.

3. Comparison between before PCA and after PCA

The following table is the result for SHG features before PCA and after PCA. The overall result before PCA is a little bit better than the result after PCA. Because some information is lost during principal components analysis. But correlation between features is eliminated. So the result is truer.

Table 27. Experimental Results Comparison between before PCA and after PCA

Features Users	SHG (Before PCA)	SHG (After PCA)
2867	79.71% J48	75% Naïve Bayes
3641	58.33% J48	66.67% Logistic
5055	88.89% Naïve Bayes	88.24% J48
Cross-Users	75.68% J48	68.18% J48

4. Comparison between Statistical Pipeline and Deep Learning Pipeline

The following table is the comparison between statistical pipeline and deep learning pipeline. The overall result of statistical pipeline is much better than deep learning pipeline. For statistical pipeline, within-user result is better than cross-user result, but it is opposite for deep learning pipeline.

Table 28. Result Comparison between Deep Learning and Statistical Pipeline

Features / Users	Skin Temp Features	Heart Rate Features	GSR Features
Within-User Statistical	70.33%	79.14%	76.51%
Within-User Deep	48.31%	54.93%	61.53%
Cross-Users Statistical	63.06%	75.68%	60.36%
Cross-Users Deep	58.59%	55.86%	63.96%

6. Conclusion

Based on the real physiological data, this thesis starts from drinking records prediction to drinking episode prediction based on statistical pipeline, then to drinking episode prediction based on deep learning pipeline. So this is a very solid work.

Unlike other related work in the lab only did prediction based on each record, this work did drinking prediction based on both each records and drinking episodes. The advantage of prediction based on episodes is that there is no overlapping information between each episode, which does not apply to prediction on each record.

Moreover, different users have different best models. And even for the same user, different features have different best models. Based on the results on single one dimensional feature, heart rate is the most significant feature for drinking prediction because it has the highest accuracy. The best result has accuracy up to 89%, which is very good for the physiological data we have.

Although a lot of work has been done in this thesis, there may be still some future work that can be done later. They will be introduced in next chapter.

7. Future Work

A lot of new features are extracted, multiple machine learning methods, as well as Cifar 10 deep learning model, are applied in my thesis work. And the best result is near 89%. But some improvement may still be made. Based on this work, there are three main things that can be done by future work.

The first thing is to take the amount of alcohol the users have drunk into account. Because currently the drinking episode is treated as thirty minutes before the drinking time and two hours after the drinking time for every situation. But the fact is that the amount of alcohol the user drinks for each time may be different and different amount of alcohol may have different length of effect on the user. So it may be more reasonable to label the drinking episode according to the amount of alcohol the user has drunk. Then the labeling will be more accurate and the experimental result should be better.

The second thing is to try more deep learning models to extract features from the spectrum graph. Then compare them to see which one can most deeply learn the potential information from the data.

The third thing is that the number of positive thirty minutes data block is too small for classification. User 2867 has the most positive data blocks, which is only 101. So the methods in this work can be applied to other drinking data in the lab.

8. References

[1]. Peng, Sun, et al. "ADA - Automatic Detection of Alcohol Usage for a Mobile Ambulatory Assessment System" The 2nd IEEE International Conference on Smart Computing (SMARTCOMP 2016), May 2016.

[2]. Shi, Ruiqi, et al. "mAAS--A Mobile Ambulatory Assessment System for Alcohol Craving Studies." Computer Software and Applications Conference (COMPSAC), 2015 IEEE 39th Annual. Vol. 3. IEEE, 2015.

[3]. Zhang, Chen. “Wearable Sensing Analysis – Identifying alcohol Drinking From Daily Physiological Data” 2016.

[4]. Wergeles, Nickolas M. “AMD: Analysis of Mood Dysregulation A Machine Learning Approach” 2016.

[5]. 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.

[6]. Wikipedia, "CIFAR-10" Wikipedia Foundation, Inc., 2017. [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html. [Accessed April 2017].

[7]. Alex Krizhevsky. “Multiple Layers of Features from Tiny Images” 2009.

[8]. Wikipedia, "Tensor Flow" Wikipedia Foundation, Inc., 2017. [Online]. Available: https://www.tensorflow.org/tutorials/deep_cnn. [Accessed April 2017].

[9]. Wikipedia, "Convolutional Neural Network" Wikipedia Foundation, Inc., 2017. [Online]. Available: https://en.wikipedia.org/wiki/Convolutional_neural_network. [Accessed April 2017].

[10]. Korekado, Keisuke; Morie, Takashi; Nomura, Osamu; Ando, Hiroshi; Nakano Teppei; Matsugu, Masakazu; Iwata, Atsushi (2003). “A Convolutional Neural Network VLSI for Image Recognition Using Merged/Mixed Analog-Digital Architecture”. Knowledge-Based Intelligent Information and Engineering Systems: 169-176. CiteSeerX 10.1.1.125.3812.

[11]. Matusugu, Masakazu; Katsuhiko Mori; Yusuke Mitari; Yuji Kaneda (2003). “Subject independent facial expression recognition with robust face detection using a convolutional neural network”. Neural Networks. 16 (5): 555-559. doi: 10. 10 16/S0893-6080(03)00115-1. Retrieved April 2017.

[12]. G. Berntson, K. Quigley, J. Jang, and S. Boysen, “An approach to artifact identification: application to heart period data,” Psychophysiology, vol. 27, no. 5, pp. 586–598, 1990.

[13]. R. Kloner, S. Hale, K. Alker, and S. Rezkalla, “The effects of acute and chronic cocaine use on the heart,” Circulation, vol. 85, no. 2, pp. 407–419, 1992.