Statistical Learning for Biological and Information Systems Final Project

Chenye Yang, Bingzhuo Wang, Zhuoyue Xing
cy2540@columbia.edu, bw2632@columbia.edu, zx2269@columbia.edu

December 19, 2019

Data Set: WISDM Smartphone and Smartwatch Activity and Biometrics Data Set
Paper: Smartphone and Smartwatch-Based Biometrics Using Activities of Daily Living

1 Introduction

1.1 Background and Application Area

Biometric identifiers are unique characteristics of people and they are used to distinguish an individual [1]. With biometric identifiers, a person can be identified or authenticated through biometric methods. Because the biometric identifiers are unique physical or behavioral characteristics of people and vary from person to person, there's no need to worry about them being copied or stolen [2], unlike the traditional security method though password or ID card [3].

Biometric methods can be divided into physiological methods and behavioral methods, depending on the identifiers they use [4, 5]. Physical biometric methods use physiological identifiers like fingerprints, face, iris and DNA [6], among which the most commonly used is fingerprints [5]. However, fingerprints may be not accurate, with Equal Error Rate (EER) value being $2\% - 7\%$ [7]. Moreover, sometimes it is difficult to use physical biometric methods [8]. Behavioral biometric methods use behavioral identifiers like walking, handwaving, signature and voice [9]. These behavioral identifiers are highly related to people's daily life and can be measured by common devices like smartphone and smartwatch, which are the devices used in the paper we chosen [8].

Gait is one of the most popular identifiers in behavioral biometric methods [10]. Gait biometrics can be implemented in three ways using different types of sensors: vision-based, floor-based, and wearable sensor-based. However, vision-based and floor-based systems is only suitable for fixed place or instrumented area [8]. Wearable sensor-based systems are not limited to specific locations, nevertheless, the sensors may cause uncomfortable for subjects. Using the most common devices, such as smartphone and smartwatch, as the sensors is more acceptable to subjects, and has being well studied. Some researches only utilized smartphone as data source [11–14], while others used smartphone and smart band or some sensors to simulate smartphone and smart band to do gait biometric authentication or identification [15–17].

Not only is gait used in behavioral biometric methods, but also "soft touchscreen" gestures like

tap and drag, "soft keystroke" gestures like duration and pressure along with accelerometer and ayroscope, "sound" properties are used to do biometric authentication with the help of only smartphone [18–22]. Moreover, the gestures when people writing specific words or their signatures are used to do identification with the help of smartwatch accelerometer and gyroscope sensors [23].

The behavioral biometric methods with more than one identifiers have been studies [24, 25]. However, they only combined a few identifiers, such as walking, standing, jogging, sitting and running, and only used smartphone as the measurement device [8].

1.2 Problems Considered

It is important to evaluate a great number of behavioral biometric identifiers which are from daily life through smartphone or smartwatch [8]. The first reason is that this process is vital to find a base behavioral identifier which can be used individually to build a biometric system for identification or authentication. The second reason is that this process may be helpful to build a biometric system which uses easy-to-get unconscious normal daily behavioral identifiers to do identification or authentication. These behavioral biometric systems, because the data is collected by daily electronic devices, can be used to identify or authenticate individuals to secure smartphone or smartwatch. Moreover, the biometric systems are also available to secure other devices like car and home, using smartphone or smartwatch as a data-collecting and data-transmitting node.

Driven by the importance of evaluation of daily identifiers, the paper we choose [8] analyzed 18 normal daily activities for behavioral biometric methods, and used the accelerometer and gyroscope sensors on smartphone and smartwatch to do identification and authentication. The paper is unique because it used a lot of daily activities and used both smartphone and smartwatch.

The paper we choose mainly focused on the following questions to conduct research:

1. The value of smartphone, smartwatch and their combination for behavioral biometric methods.
2. The choose of accelerometer and gyroscope sensors on smartphone and smartwatch.
3. The performance of behavioral biometric methods for identification and authentication.
4. The performance of daily behavioral identifiers for biometric system.
5. The amount of training data.

1.3 Report Organization

This final report is organized as follows. Section 2 describes the data set we choose and the paper using that data set. In Section 3, the results in the paper we choose are reproduced. Section 4 then compares the different statistical learning techniques used to improve the paper. Section 5 shows the results with different techniques and discusses the advantages and disadvantages of these techniques. In section 5 we also list the future work.

2 Data Set and Paper

2.1 Data Set in Detail

2.1.1 Raw Data from Sensors

The "WISDM Smartphone and Smartwatch Activity and Biometrics Data Set" is collected by Wireless Sensor Data Mining (WISDM) Lab in the Department of Computer and Information Science of Fordham University. It is collected from 51 subjects performing 18 different daily life activities. The devices used to collect the data are accelerometer and gyroscope sensors on smartphone (in subjects' pocket) and smartwatch (on subjects' dominant wrists). The sensor data are collected every 50ms (20Hz) theoretically. The general information of the data set is as follow Table 1, which can also be found in the description of the WISDM dataset on UCI Repository.The activities and their corresponding codes are shown in Table 2.

Table 1: Summary Information about Data Set [26]

Attribute	Value
Number of subjects	51
Number of activities	18
Minutes collected per activity	3
Sensor polling rate	20Hz
Smartphone used	Google Nexus 5/5x or Samsung Galaxy S5
Smartwatch used	LG G Watch
Number raw measurements	15,630,426

Table 2: Activities and Their Codes [26]

Activity	Code	Activity	Code	Activity	Code
Walking	A	Brushing Teeth	G	Kicking (Soccer Ball)	M
Jogging	B	Eating Soup	H	Playing Catch w/Tennis Ball	O
Stairs	C	Eating Chips	I	Dribbling (Basketball)	P
Sitting	D	Eating Pasta	J	Writing	Q
Standing	E	Drinking from Cup	K	Clapping	R
Typing	F	Eating Sandwich	L	Folding Clothes	S

The raw data recorded by accelerometer and gyroscope sensors are time-series data. The sampling rate of the sensors is 20Hz, which is suggested by operating systems and can be delayed by other CPU process. Sensor data is stored in text file with name like

data_1609_accel_watch.txt

data_1619_gyro_phone.txt

where *data* is a common header, *1609* is the ID of subject (from *1600* to *1650*), *accel* means the data is sampled by accelerometer sensor, *gyro* means the data is sampled by gyroscope sensor, *watch* means the sensor is on smartwatch, *phone* means the sensor is on smartphone. All the

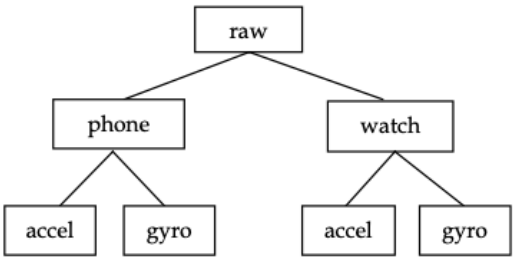

Figure 1: Data Directory [26]

sensor data are stored in such directory as Figure 1.

Within each sensor data file with the name above, each line is one sample from the sensor. The format of the raw sensor sample data is *Subject ID, Activity Code, Timestamp, x, y, z*, of which the meaning are listed in Table 3. Following are some actual data in file *data_1608_accel_phone.txt*. Please note that the units of *x, y, z* are different for different sensor types. The units of accelerometer sensor are m/s^2, while those of gyroscope sensor are radians/s.

1608,A,111545844427476,4.319275,9.2408905,-0.77568054;
1608,A,111545864569077,1.0476074,6.8250427,1.1964569;
1608,A,111545884710679,-0.78904724,5.8375854,3.246933;

Table 3: Definition of Elements in Raw Data Measurements [26]

Field Name	Description
Subject ID	Integer. Uniquely identifies the subject. Range: 1600-1650.
Activity Code	Letter. Identifies a specific activity listed in Table 2. Range: A-S (no "N")
Timestamp	Integer. Unix time
x/y/z	Real. Sensor value for x/y/z axis. May be positive or negative.

Figure 2 shows an example of the smartphone accelerometer data for walking and jogging. The *y*-axis means the vertical direction and its data has the largest magnitude, which is in line with the facts of walking and jogging. Also, the jogging activity has a higher frequency than walking.

For each raw data text file with the former name, there should be $18 \times 3 \times 60 \times 20 = 64800$ lines with sensor samples. Because four sensors were used and 51 subjects participated in data collecting, there should be $64800 \times 4 \times 51 = 13219200$ samples in total, with each sensor having 3304800 samples. However, because the collecting is not perfect, the actual lines of data samples belonging to each sensor are:

raw/phone/accel:	4804403
raw/phone/gyro:	3608635
raw/watch/accel:	3777046
raw/watch/gyro:	3440342

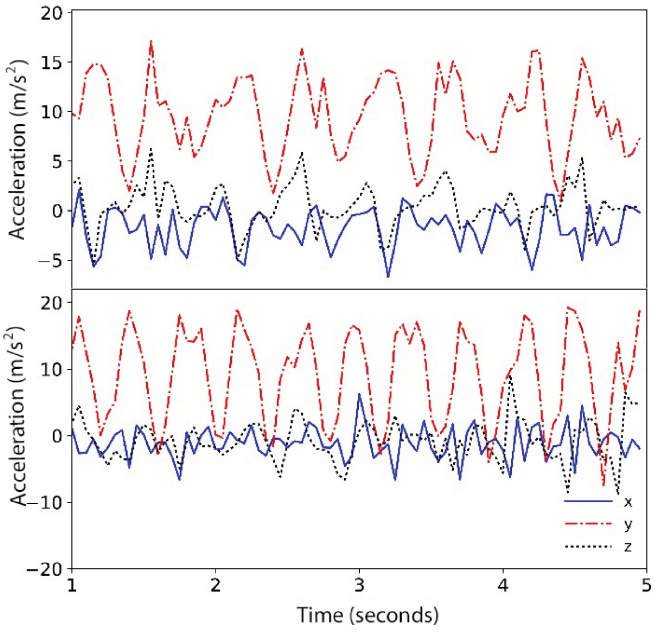

Figure 2: Graphical plot of the smartphone's triaxial accelerometer data for the walking activity (top) and the jogging activity (bottom) [8]

2.1.2 Feature Extraction

In Section 2.1.1, we get raw time-series data from accelerometer and gyroscope sensors in smartphone and smartwatch. However, many statistical learning techniques can not deal with time-series data directly. Therefore, the data need to be transformed into labeled examples with features and labels.

Considering the activities are repeatedly performed by subjects, the authors [8] choose non-overlapping 10 seconds as a segment. The 10 seconds raw data is long enough to contain information for repeated daily activities like walking and jogging, which is also proved by Figure 2. For every 10 seconds of every activities performed by every subjects, 92 Features are extracted using the 10 seconds raw data (200 pieces of raw data) and stored in ARFF (Attribute-Relation File Format) files. The 92 features are shown in the Table 4. However, in the paper, the authors only use 43 out of 92 features, which are highlighted in bold.

2.2 What was Done in the Paper

The basic task for this paper is to authenticate and identify the person to maintain the security of assets.

The authentication part aims to distinguish an authorized subject from other unauthorized subjects. Therefore, authentication is a classification problem with two classes. According to the data, there are 51 subjects in total and each subject needs a distinct model for authorization. For every subject, there should be 18 activities to evaluate, 9 sensor combinations to test and 3 classification algorithms to perform. Therefore, there should be in total 24786 experiments to execute.

Table 4: Definition of Elements in Arff Data [26]

Features or Labels	Description
ACTIVITY	The activity performed using one of activity codes from Table 2
X{0-9}, Y{0-9}, Z{0-9}	The distribution of values over the x, y, and z axes
{X,Y,Z}AVG	Average sensor value over the window
{X,Y,Z}PEAK	Time in milliseconds between the peaks in the wave associated with most activities
{X,Y,Z}ABSOLDEV	Average absolute difference between each of the 200 readings and the mean of those values
{X,Y,Z}STANDDEV	Standard deviation of the 200 values
{X,Y,Z}VAR	Variance of the values
XMFCC{0-12}, YMFCC{0-12}, ZMFCC{0-12}	MFCCs represent short-term power spectrum of a wave
{XY, XZ, YZ}COS	The cosine distrances between sensor values for pairs of axes
{XY, XZ, YZ}COR	The correlation between sensor values for pairs of axes
RESULTANT	Average resultant value, computed by squaring each matching x, y, and z value, summing them, taking the square root, and then averaging these values over the 200 readings
Class	Subject ID

To fully develop the topic, the model needs both data from the exact subject as positive samples and data from other subjects as negative samples. It is reasonable to combine all negative samples into a single class for authentication task, thus the problem can be developed to a binary classification. The training set is composed of 90 seconds each activity each sensor combination from authorized subject and 270 seconds each activity each sensor combination from unauthorized subjects, which means the rate is 1:3. Because in real life, data from unauthorized subjects are unavailable, thus the unauthorized data in test set should not overlap with that in training set. Equal Error Rate (EER) or Accuracy is used to evaluate the performance of these biometric behavioral identifiers and the combinations of sensors.

The identification part aims to assign all subjects into different classes. Therefore, identification is a classification problem with fifty-one classes. The main difference from the authentication problem is to handle a multi-class task, while other configuration should be quite similar to the authentication one. Except that, for the training set and test set, 10-fold cross validation is used to split the original data set, and training set must have information from all subjects.

2.3 Statistical Learning Tools Used in Paper

With data described above, the paper used classification algorithm to generate separate authentication and identification models to evaluate whether a person can be distinguished from others. The authors of the paper used three methods to perform the classification: K Nearest Neighbors, Decision Tree and Random Forest. These methods are quite interpretable. For K-NN, the number of neighbors is set to 5 and the distance metric is Minkowski distance metric. For Random Forest, the maximum number of features is the square root of the features of processed data and the

number of decision trees is set to 10. In the paper, other parameters of these methods, which are not specified, are preferable to the default values.

2.3.1 K-NN

K nearst neighbors [27] is based on the idea that similar argument values should lead to similar function values. In classification K-NN, the value function is a class membership. The class membership is selected by majority vote from its neighbors, which means the subject is assigned to the class most common among its K nearest neighbors. If K = 1, then the subject is assigned to the class of its single nearest neighbor.

2.3.2 Decision Tree

The idea of classification decision tree [27] is to segment the predictor space $(X_1, ..., X_p)$ into distinct and non-overlapping regions $R_1, ..., R_j$, The classification is based on majority vote over segments. The object of region division is to minimize the error classified data. In other words, minimize the following measures:

$$\text{Classification Error Rate: } E = 1 - \max_k \hat{p}_{m,k}$$

$$\text{Gini Index: } G = \sum_{k=1}^{K} \hat{p}_{m,k}\,(1 - \hat{p}_{m,k})$$

$$\text{Entropy: } D = -\sum_{k=1}^{K} \hat{p}_{m,k} \log \hat{p}_{m,k}$$

where $\hat{p}_{m,k}$, the proportion of training observations in the m-th box that are from class k.

2.3.3 Random Forest

The idea of random forest [27] is to build lots of decorrelated trees through bootstrap set. When spliting the regions in one tree, only a subset of the predictors is considered. The number of the input predictors is often chosen as the square root of the number of the original predictors. For classification usage, the decision is the majority vote result among all the classifiers (trees in forest).

3 Reproduce the Results from the Paper

3.1 Authentication

The paper provides three approaches to classification, K Nearest Neighbor(kNN), Decision Tree(DT) and Random Forest(RF) and shows the accuracy results of the best algorithm (Random Forest) with 9 sensor combinations per activity and in average.

To choose the best size of train data per activity, we run the three algorithms on different size of data. Figure 3 shows the accuracy averaged all eighteen activities versus the amount of train data per activity.

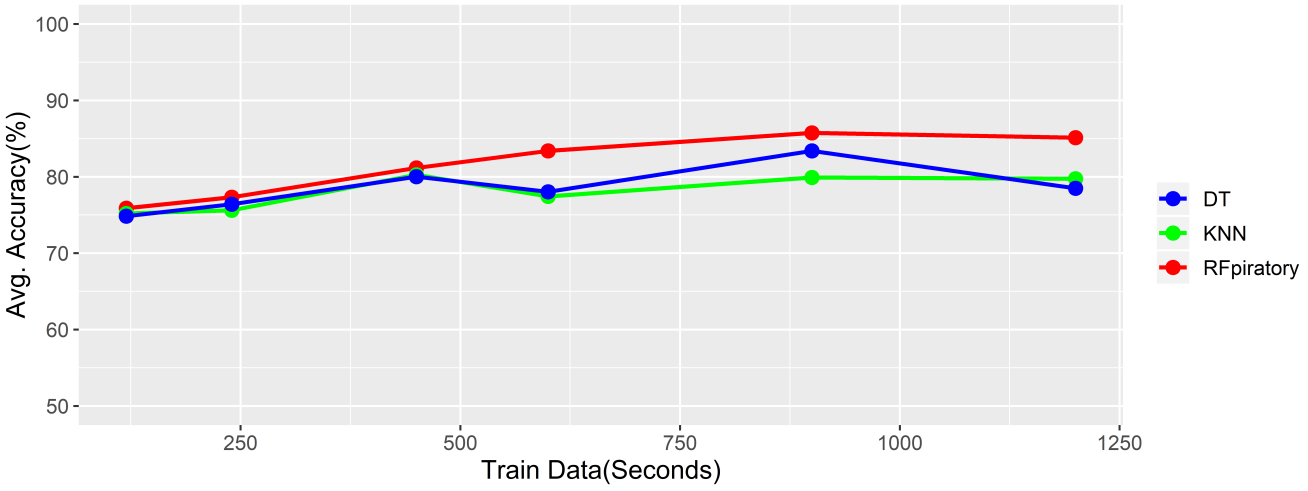

Figure 3: Graphical plot of the accuracy averaged all eighteen activities versus the amount of train data per activity

It can be found that when the size of train data per activity increases, the accuracy is not strictly increasing. On the contrary, the accuracy sometimes decrease. On reason for this condition is we select train data and test data randomly from our data set. Another reason is our model has too many predictors, leading our model sensitive to variance and easier to overfit train data. Thus, we choose the size of train data per activity as 900 seconds.

We reproduce results of all three algorithms to compare their accuracy and find the best algorithm to do authentication. Within every algorithm, the authentication process varies in activities and used sensors, which can be a sensor or sensors combination. In view of that, we compute the classification accuracy with every single activity and every type of used sensors. To briefly describe the accuracy and make comparison between different types of used sensors, we compute the average accuracy among all activities, which is shown in the Table 5. Each row of Table 5 is the average result on the accuracy result of different activities in the paper.

Table 5: Authentication accuracy results

Algorithm	Accel Phone	Gyro Phone	Accel Watch	Gyro Watch	Phone	Watch	Accel	Gyro	All
RF	**0.92**	0.85	0.87	0.80	**0.88**	0.84	**0.88**	0.81	0.87
DT	0.88	0.78	0.78	0.75	0.82	0.75	0.81	0.76	0.78
kNN	0.87	0.78	0.80	0.74	0.82	0.77	0.85	0.76	0.80

From Table 5, Random Forest has the best performance on the authentication problem. At the same time, using only the data from the accelerometer on phone has the best performance on the authentication problem. The paper also believe the Random Forest algorithm is the best algorithm. However, the result paper can not show that using only the data from the accelerometer on phone is the best choice. The reason of this difference between the paper and our results may be the difference on data. Because there are some duplicate data in the dataset we download from Internet. The author may upload these data by mistake. Table 6, Table 7 and Table 8 show

the average accuracy on all the person for every activities and every type of data. In Table 8, it can be observed that using the C(Stairs), L(Eating Sandwich), Q(Writing) activity data from accelerometer on phone to do subject authentication is the best choice.

Table 6: K-NN Result for Authentication

Activity Code	Accel Phone	Gyro Phone	Accel Watch	Gyro Watch	Phone	Watch	Accel	Gyro	All
A	80.8	80.0	80.8	71.9	80.0	74.7	82.2	76.1	81.3
B	84.4	77.8	81.1	74.2	83.6	79.0	88.8	75.3	81.0
C	91.1	79.4	79.7	76.9	81.5	79.6	86.4	75.0	80.4
D	82.8	83.9	83.1	76.1	79.6	74.4	85.4	75.1	81.8
E	82.5	79.7	84.2	78.3	85.0	73.5	89.0	76.9	78.5
F	86.4	74.7	79.2	70.8	81.0	77.5	81.0	76.8	78.5
G	93.1	70.6	82.2	72.5	79.6	78.2	81.8	75.8	78.8
H	77.2	76.1	83.3	71.4	81.1	76.4	84.0	77.1	82.5
I	81.4	75.8	78.9	67.5	81.1	77.6	82.2	73.8	80.5
G	81.4	82.8	78.1	74.7	82.2	78.6	82.1	74.2	80.8
K	93.1	76.7	75.6	76.9	72.9	74.2	80.8	77.2	80.8
L	93.1	78.3	78.9	75.3	85.0	78.1	86.8	77.6	82.5
M	90.3	76.1	85.6	77.2	84.6	76.4	84.7	76.1	80.0
O	83.9	79.4	80.6	75.0	80.8	74.6	80.8	77.9	78.6
P	95.3	78.3	79.7	72.8	82.9	78.1	86.9	73.1	79.0
Q	93.9	70.0	79.2	75.6	86.7	77.8	85.8	73.8	79.4
R	94.2	82.2	77.8	76.9	77.2	79.9	85.6	75.4	79.7
S	89.2	73.6	81.1	70.6	80.8	78.3	86.3	77.1	78.3

In Table 8, if we choose L(Eating Sandwich) data from accelerometer on phone, we will have 99.2% accuracy on authentication. However, eating sandwich is not a good activity for authentication. We can not ask users to eat a sandwich every time we need to get the result of authentication. Instead, it would be better to choose R(Clapping) data from accelerometer on phone, because it is much more easier to clap than eat a sandwich and have an acceptable accuracy 96.7%.

Table 7: Decision Tree Result for Authentication

Activity Code	Accel Phone	Gyro Phone	Accel Watch	Gyro Watch	Phone	Watch	Accel	Gyro	All
A	82.8	75.0	80.3	75.0	81.9	75.0	78.1	75.0	77.6
B	85.3	81.7	78.9	75.0	79.2	73.5	79.9	76.8	77.8
C	90.6	72.8	75.0	75.0	86.0	75.0	82.5	80.0	77.6
D	78.1	78.1	82.8	75.0	75.8	75.0	79.2	75.0	76.7
E	81.9	76.1	81.1	75.0	85.1	75.0	80.8	75.0	78.3
F	87.2	79.2	77.5	75.0	81.4	75.0	81.4	75.0	73.8
G	96.1	78.6	80.0	75.0	81.1	75.0	79.2	76.5	77.4
H	83.1	75.0	78.6	75.0	85.4	75.0	82.5	75.0	79.0
I	87.8	78.3	75.8	75.0	81.9	78.3	80.6	75.0	78.2
G	84.4	79.7	69.7	75.0	82.4	75.0	82.2	75.6	76.8
K	96.7	71.7	78.6	75.0	81.8	75.3	76.9	76.0	77.3
L	83.3	80.8	76.1	75.0	81.1	75.0	78.1	75.0	77.6
M	85.3	76.7	77.8	75.0	81.1	75.0	80.0	75.0	76.7
O	89.7	77.5	75.0	75.0	78.2	75.0	82.9	75.0	78.8
P	91.4	84.4	79.4	75.0	81.8	75.0	78.8	75.0	76.7
Q	92.5	72.5	80.6	75.0	78.9	75.0	84.7	75.0	78.9
R	96.1	83.1	76.9	75.0	82.5	75.0	79.9	75.0	76.7
S	89.2	81.4	79.2	75.0	82.6	74.3	82.6	79.3	80.6

Table 8: Random Forest Result for Authentication

Activity Code	Accel Phone	Gyro Phone	Accel Watch	Gyro Watch	Phone	Watch	Accel	Gyro	All
A	86.1	89.2	87.5	75.0	82.5	84.9	87.5	84.6	86.0
B	92.8	83.1	84.4	79.2	89.7	82.8	91.8	82.9	87.9
C	**97.8**	83.1	88.9	85.3	90.6	83.1	88.8	80.4	85.2
D	77.5	85.8	86.1	83.1	85.6	84.7	89.2	77.9	85.2
E	86.7	84.2	88.6	82.2	88.9	80.0	92.2	82.5	87.7
F	91.7	84.2	89.7	82.8	91.3	84.2	80.0	80.0	85.0
G	96.4	78.9	90.3	77.5	90.1	84.3	86.9	77.5	87.7
H	90.3	88.3	87.2	80.3	88.8	82.8	87.1	80.0	87.3
I	90.8	86.7	84.7	82.8	88.9	82.8	89.3	80.0	86.9
G	85.8	84.7	83.6	83.1	87.2	86.4	85.3	79.2	86.3
K	96.1	83.1	81.7	80.8	84.0	85.8	86.9	80.0	89.4
L	**99.2**	84.2	85.0	75.0	91.9	83.1	92.5	79.6	86.5
M	91.1	86.9	89.2	83.3	87.8	86.5	87.9	79.2	87.1
O	91.1	86.4	86.9	80.6	84.4	83.6	86.7	77.9	86.3
P	96.7	86.9	87.2	82.5	91.8	83.2	86.0	80.0	84.0
Q	**97.2**	78.3	88.6	80.6	92.8	83.1	90.7	85.8	86.0
R	**96.7**	85.6	81.7	75.0	83.9	83.6	92.1	80.0	86.3
S	95.6	86.7	90.0	77.5	89.0	83.8	90.7	82.5	86.5

3.2 Identification

Subject identification requires a classifier to give distinguishable label assigned to different subjects. In this section, we will cover all three methods mentioned in the paper and reproduce the corresponding results separately. For a brief view, Table 9 shows the results from K Nearest Neighbor method, and Table 10 for Tree-based method, Table 11 for Random Forest method.

Within every method, the identification process varies in activities and sensor combinations. In view of that, we compute the classification accuracy with every single activity and every sensor combination. To briefly describe the accuracy and make comparison between different sensor combinations, we compute the average accuracy among all activities.

Table 9: K-NN Result for Identification

Activity	Accel Phone	Gyro Phone	Accel Watch	Gyro Watch	Phone	Watch	Accel	Gyro	All
Walking	88.6	51.3	49.1	29.7	94.2	59.1	89.5	59.1	93.6
Jogging	88.1	63.8	59.2	33.1	90.3	68.0	85.9	74.3	89.1
Stairs	72.7	25.6	30.2	21.2	77.5	37.9	70.2	45.3	78.6
Sitting	87.7	30.7	60.4	22.4	76.2	55.8	85.6	44.4	74.0
Standing	70.4	24.4	47.3	18.8	60.4	44.7	76.6	38.3	62.6
Kicking	87.1	28.0	59.6	18.5	84.4	62.1	91.0	46.6	83.6
Dribbling	81.9	24.5	51.4	24.3	79.3	58.8	85.4	46.4	82.2
Catch	86.3	20.4	50.4	22.0	73.9	57.8	88.5	41.0	72.7
Typing	82.9	24.5	52.8	16.7	75.8	52.8	89.5	42.1	78.3
Writing	84.8	17.9	51.9	13.7	72.4	48.9	89.1	32.2	73.0
Clapping	87.7	20.8	52.6	18.5	73.6	53.2	89.5	36.0	74.6
Teeth	85.8	24.1	49.8	12.0	75.7	45.9	88.3	39.3	77.5
Folding	73.2	20.1	26.7	14.4	72.9	32.2	68.8	37.0	72.8
Pasta	80.0	26.5	51.3	21.2	77.0	49.9	82.8	47.0	76.3
Soup	79.7	27.6	55.6	31.6	80.7	61.3	88.0	59.3	80.5
Sandwich	88.1	30.5	58.1	16.8	85.6	55.4	90.4	41.6	85.1
Chips	87.2	25.1	79.3	43.2	83.3	83.3	94.1	57.9	83.3
Drinking	80.3	24.6	42.2	13.8	78.0	40.2	80.6	39.3	78.3
Avg	**82.9**	28.4	51.5	21.8	78.4	53.7	**85.2**	45.9	78.7

As the Table 9, Table 10 and Table 11 indicate, the accuracy using accelerometer on smartphone performs quite good results using various methods, it is quite convincing that the data from the accelerometer on smartphone can lead to better results, besides, the accelerometer on watch can also provide some hints on predicting which subject the activity belongs to, it may not be straightforward to distinguish subjects by using single data, but the combination of the watch and phone leads to a better result, which shows that data from different sensors or sensor combinations may enhance the model performance although it varies.

For method evaluation, the basic idea is that random forest provides better results among all sets of data although training time costs more correspondingly. Since the results varies from both different models and parameters, the default value used in the paper gives brief idea about the

Table 10: Decision Tree Result for Identification

Activity	Accel Phone	Gyro Phone	Accel Watch	Gyro Watch	Phone	Watch	Accel	Gyro	All
Walking	92.9	87.3	66.6	53.9	91.9	69.3	89.9	86.0	91.7
Jogging	89.7	83.4	62.8	63.3	88.8	70.3	87.6	82.3	89.4
Stairs	84.8	60.3	44.8	30.6	85.5	46.1	83.2	61.1	85.1
Sitting	94.1	49.6	75.6	27.7	91.1	72.1	93.0	45.4	91.6
Standing	92.4	38.3	71.4	22.4	89.3	68.4	91.2	37.8	90.3
Kicking	96.8	58.3	79.3	38.3	94.8	82.6	94.3	60.5	94.6
Dribbling	95.7	58.9	62.0	48.9	92.0	60.8	92.7	65.7	93.4
Catch	95.6	46.4	72.2	43.4	93.2	73.0	95.4	51.8	93.7
Typing	96.4	43.1	59.8	26.8	94.5	61.2	92.6	47.3	94.0
Writing	93.3	42.2	67.0	29.6	94.3	66.9	93.5	44.5	94.0
Clapping	95.5	47.3	67.7	29.2	93.8	65.5	93.3	51.1	93.6
Teeth	95.9	48.1	57.3	24.1	93.4	55.7	94.3	44.6	93.6
Folding	82.2	43.4	43.8	25.0	80.2	47.1	83.9	46.9	81.1
Pasta	87.8	43.8	54.0	50.5	86.1	59.8	89.2	54.4	86.9
Soup	92.5	47.2	63.1	66.3	89.1	72.5	89.0	69.9	90.9
Sandwich	93.9	60.5	78.3	42.0	90.9	78.1	92.2	58.1	91.3
Chips	95.3	60.5	77.1	68.6	92.4	76.6	93.6	75.8	93.6
Drinking	89.2	45.9	44.4	23.3	85.8	50.3	86.3	49.7	85.0
Avg	**92.4**	53.6	63.7	39.7	90.4	65.4	**90.8**	57.4	**90.8**

performance of different model. K nearest neighbor is an easy way to classify the data with acceptable interpretability, but it can be quite unstable for certain field of data. Decision tree is also time saving and with high interpretability, yet the various metrics may provide with totally different results. Random forest uses ensemble model to enhance general performance, however, the time cost rises along with the complexity of model, which requires better computational support.

Table 11: Random Forest Result for Identification

Activity	Accel Phone	Gyro Phone	Accel Watch	Gyro Watch	Phone	Watch	Accel	Gyro	All
Walking	98.5	96.1	86.3	73.1	98.9	88.6	98.8	98.9	98.8
Jogging	95.9	93.0	86.3	85.5	95.8	92.0	96.1	95.5	95.1
Stairs	93.9	83.1	71.6	49.7	94.6	78.3	95.3	94.9	94.5
Sitting	95.9	64.9	80.8	47.9	94.2	80.5	95.0	94.1	94.0
Standing	94.0	58.1	81.0	36.2	94.1	79.0	96.3	94.3	94.1
Kicking	98.3	77.5	89.3	59.3	97.3	92.2	98.2	97.7	97.6
Dribbling	96.9	77.3	81.9	63.1	95.9	81.9	96.7	96.2	96.2
Catch	96.9	64.6	86.8	64.8	96.7	90.0	97.8	96.6	96.4
Typing	97.6	65.1	80.6	47.7	96.7	84.4	97.7	96.8	96.2
Writing	97.2	64.5	84.1	50.9	97.3	87.5	98.1	97.1	97.3
Clapping	97.6	67.2	81.7	49.2	96.8	84.3	97.6	96.4	96.4
Teeth	97.1	67.0	78.9	41.6	97.3	80.1	98.0	97.0	97.0
Folding	92.2	68.3	64.1	43.7	92.9	72.1	93.8	93.9	93.7
Pasta	96.1	69.4	82.3	81.6	96.2	88.9	97.8	95.7	95.6
Soup	96.9	68.2	87.0	86.6	97.1	92.6	98.4	97.2	96.9
Sandwich	95.9	78.9	88.2	66.0	94.0	89.4	94.6	93.9	94.3
Chips	97.3	76.9	90.3	86.1	96.0	93.6	97.6	96.2	96.3
Drinking	94.0	68.8	70.7	49.8	94.3	80.3	96.0	93.4	93.9
Avg	**96.2**	72.7	81.8	60.1	95.9	85.3	**96.9**	95.9	95.8

4 Different Techniques

4.1 Support Vector Machine

The idea of support vector machine [27] is to separate data points using hyperplanes. The choice of the separate surface should maximum the minimal margin, the distance from an observation to the hyperplane. If we allow soft nonlinear classifier, the hyperplane is the solution to:

$$\max_{\beta_j, \epsilon_j M} M$$
$$\text{s.t.} \begin{cases} \sum_{1}^{p} \sum_{k=1}^{2} \beta_{jk}^2 = 1 \\ y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_{j1} x_{ij} + \sum_{j=1}^{p} \beta_{j2} x_{ij}^2 \right) \geq M \left(1 - \epsilon_i\right), \quad \forall i \\ \epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C \end{cases}$$

The support vector classifier can be represented as:

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}}^{n} \alpha_i \langle x, x_i \rangle$$

where α_i are training parameters and $\mathcal{S}$ is the set of support points.

Kernel $K(x_i, x_j)$ is a generalization of the inner product $\langle x_i, x_j \rangle$, which with the decision function can be written as:

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}}^{n} \alpha_i K(x, x_i)$$

Followings are some common kernels to use:

$$\begin{aligned}
&\text{Linear} && K(x_i, x_j) = \langle x_j, x_i \rangle \\
&\text{Polynomial} && K(x_i, x_j) = (1 + \langle x_j, x_i \rangle)^d \text{ for positive integer } d \\
&\text{Radial} && K(x_i, x_j) = \exp\left(-\gamma \sum_{k=1}^{p} (\langle x_{ik} - x_{jk} \rangle)^2\right) \text{ for } \gamma > 0
\end{aligned}$$

4.2 SVM for Authentication

In this part, we try to apply the Support Vector Machine method into the data collected from 51 subjects in different activities and sensor combinations. Because the scale our data set is too large, it takes too much time to run SVM algorithm whose kernel is not linear function and we choose linear kernel. The result of using linear kernel SVM are listed below in Table 12.

Table 12: Support Vector Machine Result for Authentication

Activity Code	Accel Phone	Gyro Phone	Accel Watch	Gyro Watch	Phone	Watch	Accel	Gyro	All
A	89.7	80.3	83.6	75.6	81.7	72.5	84.6	75.0	77.7
B	89.2	78.6	84.4	75.3	79.2	70.4	75.0	73.8	76.9
C	88.9	76.9	78.9	76.7	79.6	74.2	79.2	76.7	75.8
D	81.1	76.7	85.0	76.9	75.0	72.5	79.6	74.6	78.3
E	87.5	75.6	89.2	75.8	79.6	74.2	83.8	74.2	76.5
F	88.1	78.3	80.8	73.9	70.4	69.6	78.8	75.0	78.8
G	**96.9**	70.6	85.0	73.3	71.7	69.2	82.5	70.0	78.3
H	89.2	80.6	84.2	76.4	78.8	75.4	84.6	75.0	77.5
I	91.4	75.3	82.5	73.6	77.5	71.3	81.3	73.3	76.3
G	83.9	77.8	80.0	75.0	77.9	70.8	83.8	74.2	72.5
K	94.4	77.8	72.5	76.9	81.3	71.7	77.5	72.5	77.7
L	91.7	71.7	84.2	75.0	85.8	78.8	77.9	72.1	75.4
M	89.4	77.2	83.9	78.3	82.1	70.8	76.3	72.9	78.3
O	89.4	79.4	82.2	74.4	76.7	67.1	84.6	69.6	77.7
P	95.8	76.7	77.8	74.7	81.3	78.8	79.2	75.0	76.5
Q	**96.7**	65.6	84.4	75.0	82.5	69.2	85.0	75.0	77.5
R	**91.9**	71.1	83.1	75.0	79.6	72.5	78.3	70.8	80.2
S	90.8	75.3	81.9	70.0	76.3	72.5	77.5	75.0	76.7

In table 12, if we choose G(Brushing Teeth) data from accelerometer on phone, we will have 96.7% accuracy on authentication, which is the best accuracy when we use SVM with linear

kernel. However, brushing teeth could be inconvenient for some users. It is annoying that ask users to brush their teeth every time we need to get the result of authentication. Instead, it would be better to choose R(Clapping) data from accelerometer on phone, because it is much more easier to clap than brushing teeth and have an acceptable accuracy 91.9%.

4.3 SVM for Identification

In this part, we try to apply the Support Vector Machine method into the data collected from 51 subjects with different activities and sensor combinations. The result of experiments are listed below in Table 13.

Table 13: Support Vector Machine Result for Identification

Activity	Accel Phone	Gyro Phone	Accel Watch	Gyro Watch	Phone	Watch	Accel	Gyro	All
Walking	98.2	74.7	89.1	48.3	95.6	74.2	98.6	77.8	96.0
Jogging	96.3	77.9	89.6	60.0	92.8	82.8	96.6	83.0	93.0
Stairs	93.8	49.5	71.1	36.5	89.9	60.7	95.4	57.6	88.8
Sitting	94.9	39.7	70.1	25.9	88.5	68.1	92.5	47.9	89.0
Standing	89.4	34.4	68.2	23.2	85.8	63.8	89.4	46.4	85.8
Kicking	97.4	43.2	76.4	24.5	92.1	80.9	96.8	51.5	92.7
Dribbling	94.2	37.1	80.0	42.8	87.7	71.1	93.1	56.5	88.5
Catch	96.3	31.8	73.6	32.5	90.9	74.8	95.2	50.3	91.8
Typing	95.8	37.2	65.1	26.3	90.6	71.6	95.0	46.5	92.6
Writing	96.6	32.8	74.3	27.1	90.7	72.3	96.5	43.3	90.5
Clapping	96.2	34.8	70.6	22.4	91.1	70.2	95.6	43.5	91.7
Teeth	94.1	39.4	61.9	21.2	91.5	64.4	95.0	44.0	92.1
Folding	91.9	33.7	65.5	33.3	85.8	55.3	92.7	49.7	85.3
Pasta	95.3	36.9	84.0	53.4	88.5	68.9	98.1	57.5	89.3
Soup	96.3	38.3	89.2	64.5	87.6	78.6	98.3	70.3	88.9
Sandwich	93.8	44.2	81.3	23.7	90.4	74.0	92.9	49.7	90.9
Chips	95.9	37.9	90.7	56.9	93.3	85.3	97.1	66.1	93.1
Drinking	93.1	32.4	74.2	27.0	87.3	62.8	93.3	47.7	87.0
Avg	**95.0**	42.0	76.4	36.1	90.0	71.1	**95.1**	55.0	90.4

The basic idea for support vector machine in multi-class task is to divide the whole problem into binary classification sub-problems. In table 13, the average accuracy of all 18 activities in different sensor combinations varies, but it is consistent with the former data analysis steps that the single data from accelerometer on the phone and the combined data from accelerometer both on phone and watch give quite good results as 95%. However, the existence of non-ignorable variance between different sensor combinations indicates the inappropriateness of handling with multi-class data using support vector machine.

5 Discussion and Conclusion

5.1 Compare Different Techniques

Table 14 gives a brief comparison of four methods we use. Among all the four methods, Random Forest provides better performance than K-NN and Tree-based method, with the defect of less interpretability and slower training time. Support Vector Machine, which performs well in binary classification problem, does not give any better performance than Random Forest nonetheless. This is partly due to the difficulty in hot-coding process to transform binary classification into multi-class one. In general, the pros and cons (on this problem and data set) of four statistical learning methods used in this report are shown in Table 15.

Table 14: Method Comparison for Identification

Method	Accel Phone	Gyro Phone	Accel Watch	Gyro Watch	Phone	Watch	Accel	Gyro	All
K-NN	82.9	51.5	28.4	21.8	78.4	53.7	85.2	45.9	78.7
Decision Tree	92.4	63.7	53.6	39.7	90.4	65.4	90.8	57.4	90.8
Random Forest	96.2	81.8	72.7	60.1	95.9	85.3	96.9	95.9	95.8
SVM	95.0	42.0	76.4	36.1	90.0	71.1	95.1	55.0	90.4

Table 15: Pros and Cons

Method	Pros	Cons
K-NN	High interpretability Rather fast	Relatively low accuracy
Decision Tree	High interpretability Rather fast	Unstable with small change in data
Random Forest	Covariance reducing High accuracy	More time-consuming to train
SVM	-	Not easy to interpret Not good performance in multi-class classification

5.2 Future Work

1. Other kernel function can be chosen instead of linear for our SVM algorithm. It has been found that the polynomial kernel is better for this problem. However, it almost takes more than 8 hours to run a polynomial kernel SVM on one subject and more than 400 hours on the whole data set. In the future, we may switch to other program language whichever are more efficient than R to run polynomial kernel SVM.
2. The implement of decision tree method for this problem is CART. Considering the fact that decision tree is a relatively fast algorithm to solve this problem, other kinds of DT implements could be used to find a faster algorithm with acceptable accuracy.

3. There are lots of other unused statistical learning techniques such as community detection for us to deploy on the biometric behavioral method.
4. Different feature extraction methods for time-series data, besides the default way provided in data set, are also worthy to be tried.
5. Dimension reduction methods like PCA should be used to pre-produce data. Otherwise the computing process will cost a huge amount of time, with numerous extreme-high-dimension data.

References

[1] A. Jain, L. Hong, and S. Pankanti, "Biometric identification," *Communications of the ACM*, vol. 43, no. 2, pp. 90–98, 2000.

[2] A. Jain, R. Bolle, and S. Pankanti, "Introduction to biometrics," in *Biometrics*, pp. 1–41, Springer, 1996.

[3] L. O'Gorman, "Comparing passwords, tokens, and biometrics for user authentication," *Proceedings of the IEEE*, vol. 91, pp. 2021–2040, Dec 2003.

[4] K. Delac and M. Grgic, "A survey of biometric recognition methods," in *Proceedings. Elmar-2004. 46th International Symposium on Electronics in Marine*, pp. 184–193, IEEE, 2004.

[5] A. K. Jain and A. Kumar, "Biometric recognition: an overview," in *Second generation biometrics: The ethical, legal and social context*, pp. 49–79, Springer, 2012.

[6] I. Bouchrika, "A survey of using biometrics for smart visual surveillance: Gait recognition," in *Surveillance in Action*, pp. 3–23, Springer, 2018.

[7] R. Cappelli, D. Maio, D. Maltoni, J. L. Wayman, and A. K. Jain, "Performance evaluation of fingerprint verification systems," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 1, pp. 3–18, 2005.

[8] G. M. Weiss, K. Yoneda, and T. Hayajneh, "Smartphone and smartwatch-based biometrics using activities of daily living," *IEEE Access*, vol. 7, pp. 133190–133202, 2019.

[9] A. Alzubaidi and J. Kalita, "Authentication of smartphone users using behavioral biometrics," *IEEE Communications Surveys Tutorials*, vol. 18, pp. 1998–2026, thirdquarter 2016.

[10] D. Gafurov, "A survey of biometric gait recognition: Approaches, security and challenges," in *Annual Norwegian computer science conference*, pp. 19–21, Annual Norwegian Computer Science Conference Norway, 2007.

[11] M. O. Derawi, C. Nickel, P. Bours, and C. Busch, "Unobtrusive user-authentication on mobile phones using biometric gait recognition," in *2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 306–311, Oct 2010.

[12] C. Nickel, H. Brandt, and C. Busch, "Benchmarking the performance of svms and hmms for accelerometer-based biometric gait recognition," in *2011 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pp. 281–286, Dec 2011.

[13] F. Juefei-Xu, C. Bhagavatula, A. Jaech, U. Prasad, and M. Savvides, "Gait-id on the move: Pace independent human identification using cell phone accelerometer dynamics," in *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 8–15, Sep. 2012.

[14] T. Hoang, T. D. Nguyen, C. Luong, S. Do, and D. Choi, "Adaptive cross-device gait recognition using a mobile accelerometer.," *JIPS*, vol. 9, no. 2, p. 333, 2013.

[15] A. H. Johnston and G. M. Weiss, "Smartwatch-based biometric gait recognition," in *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–6, Sep. 2015.

[16] N. Al-Naffakh, N. Clarke, F. Li, and P. Haskell-Dowland, "Unobtrusive gait recognition using smartwatches," in *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–5, Sep. 2017.

[17] G. Cola, M. Avvenuti, F. Musso, and A. Vecchio, "Gait-based authentication using a wrist-worn device," in *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, MOBIQUITOUS 2016, (New York, NY, USA), pp. 208–217, ACM, 2016.

[18] T. Feng, Z. Liu, K.-A. Kwon, W. Shi, B. Carbunar, Y. Jiang, and N. Nguyen, "Continuous mobile authentication using touchscreen gestures," in *2012 IEEE Conference on Technologies for Homeland Security (HST)*, pp. 451–456, IEEE, 2012.

[19] B. Draffin, J. Zhu, and J. Zhang, "Keysens: Passive user authentication through micro-behavior modeling of soft keyboard interaction," in *Mobile Computing, Applications, and Services* (G. Memmi and U. Blanke, eds.), (Cham), pp. 184–201, Springer International Publishing, 2014.

[20] A. Buriro, B. Crispo, F. Del Frari, and K. Wrona, "Touchstroke: Smartphone user authentication based on touch-typing biometrics," in *International Conference on Image Analysis and Processing*, pp. 27–34, Springer, 2015.

[21] M. Kunz, K. Kasper, H. Reininger, M. Möbius, and J. Ohms, "Continuous speaker verification in realtime," *BIOSIG 2011–Proceedings of the Biometrics Special Interest Group*, 2011.

[22] Y. Yang, F. Hong, Y. Zhang, and Z. Guo, "Person authentication using finger snapping—a new biometric trait," in *Chinese Conference on Biometric Recognition*, pp. 765–774, Springer, 2016.

[23] F. Ciuffo and G. M. Weiss, "Smartwatch-based transcription biometrics," in *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, pp. 145–149, Oct 2017.

[24] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Cell phone-based biometric identification," in *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1–7, Sep. 2010.

[25] M. Ehatisham-ul Haq, M. A. Azam, J. Loo, K. Shuang, S. Islam, U. Naeem, and Y. Amin, "Authentication of smartphone users based on activity recognition and mobile sensing," *Sensors*, vol. 17, no. 9, p. 2043, 2017.

[26] D. Dheeru and E. K. Taniskidou, "Uci machine learning repository." `http://archive.ics.uci.edu/ml`, 2017. Accessed: 2019-12-16.

[27] P. R. Jelenkovic, *EECS E6690: Statistical Learning for Biology and Information System*. Columbia University, 2019 Fall.