Biometrics for SmartPhones: Age Recognition, Gender Recognition and Identification

A thesis submitted to the
Graduate School of Natural and Applied Sciences

by

Kamile Nur SEVİŞ

in partial fulfillment for the
degree of Master of Science

in
CyberSecurity Engineering

İSTANBUL
ŞEHİR
UNIVERSITY

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science in CyberSecurity Engineering.

APPROVED BY:

Prof. Dr. Ensar Gül ....
(Thesis Advisor)

Asst. Prof. Ali Çakmak 12/01/2017

Asst. Prof. Müjdat Soytürk 12/01/2017

This is to confirm that this thesis complies with all the standards set by the Graduate School of Natural and Applied Sciences of İstanbul Şehir University:

DATE OF APPROVAL: 12-01-2017

SEAL/SIGNATURE:

Declaration of Authorship

I, Kamile Nur SEVİŞ, declare that this thesis titled, 'Biometrics for SmartPhones: Age Recognition, Gender Recognition and Identification ' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: K. Nur

Date: 12/01/2017

"If you can't fly then run, if you can't run then walk, if you can't walk then crawl, but whatever you do you have to keep moving forward."

Martin Luther King Jr.

Biometrics for SmartPhones: Age Recognition, Gender Recognition and Identification

Kamile Nur SEVİŞ

Abstract

Identification, gender and age recognition are directly used especially in the security field and there are many researches that are actively conducted nowadays about these topics. Sensors are required for this studies. But the large and constrained sensors make it difficult for the experiments to be practicable. On the other hand, the use of smartphones are becoming increasingly popular and becoming an integral part of our lives. Therefore, the method of providing the necessary sensors via smartphones is increasing. For this study, accelerometers were used which are built-in smartphones and preferred especially to recognize motions. In a nutshell, the purpose of this thesis is to present, the study was done to recognize identification, age, and gender using the accelerometer sensor in a smartphone and the results of these studies. To accurately classify the identification, age groups, and genders that make the activities of walking and running. We also aimed to determine that which motions gives us better success rates in this classification. In addition, we aim to compare between the five different classification algorithms. For this purpose, to collect data, and to be able to use built-in sensors, compatible Android and IOS applications have been developed. Also, feature vectors were derived from the Matlab tool and classified by using the Weka Data Mining tool. Five different classifiers have been studied and the performance evaluation and comparisons were made.

Keywords: Biometric, Identification, Classification,Feature, SmartPhone, Sensor, Accelerometer, Application

Akıllı Telefonlar için Biyometri: Kimlik Doğrulama, Yaş Tanıma, Cinsiyet Tanıma

Kamile Nur SEVİŞ

Öz

Kimlik tanıma başta olmak üzere, cinsiyet tanıma ve yaş tanıma konuları özellikle güvenlik alanında doğrudan uygulamaları olan ve günümüzde aktif olarak çalışılan araştırma konularıdır. Tanımlama yapılırken gerekli olan sensor için ise hayatımızın ayrılmaz bir parçası haline gelen akıllı telefonların kullanımı da artmaktadır. Bu çalışmamızda özellikle eylem tanıma için kullanılan ve akıllı telefonlarda bulunan ivmeölçer sensörü kullanılmıştır. Özetle bu tezin amacı telefonlar üzerinde bulunan ivmeölçer sensörünü kullanarak, kimlikleri, yaş gruplarını ve cinsiyetleri tanımlamaya yonelik yapılan çalışmaları ve bu çalışmaların sonuçlarını değerlendirmektir. Yürüme ve koşma hareketlerini yapan kimlikleri, cinsiyetleri ve yaş gruplarını doğru bir şekilde sınıflandırabilmek ve yüksek başarı oranının hangi hareket ile sağlandığının tespitini yapmaktır. Bununla birlikte beş farklı sınıflandırma algoritması arasında başarım kıyaslaması yapmaktır. Bu amaçla veri toplayabilmek için, sensor olarak kullanabilecegımiz, hedeflediğimiz telefon modelleri ile uyumlu Android ve IOS uygulamaları geliştirilmiştir. Ayrıca toplanılan verilerden öznitelik vektörleri Matlab aracılığıyla elde edilmiş ve sınıflandırma Weka Data Mining Tool aracılığıyla yapılmıştır. Beş farklı sınıf üzerinde çalışma yapılmış olup, performans değerlendirmesi ve karşılaştırması hedeflenmektedir.

Anahtar Sözcükler: Biyometri, Kimlik Tanımlama, Sınıflandırma, Öznitelik, Akıllı Telefon, Sensör, İvmeölçer, Uygulama

My Dear Family and My Beloved Husband always with me...

Acknowledgments

I would like to thank you my advisor, Prof. Dr. Ensar Gül, for guiding and supporting me with all his knowledge and experience.

I also would like to thank my family for the support and constant encouragement.

Finally, I would like to thank my beloved husband for always being there for me and for his love, support and patience throughout my study.

Contents

List of Figures

List of Tables

Abbreviations

3FA	**T**hree **F**actor **A**uthentication
DNA	**D**eoxyribo **N**ucleic **A**cid
DT	**D**ecision **T**ree
GNU	**G**eneral **P**ublic **L**icense
GPRS	**G**eneral **P**acket **R**adio **S**ervice
GPS	**G**lobal **P**ositioning **S**ystem
HMM	**H**idden **M**arkov **M**odel
Hz	**H**ertz
KNN	**K** **N**earest **N**eighbors
MAP	**M**aximum **A** **P**osteriori
MATLAB	**MAT**rix **LAB**oratory
MAX	**M**aximum
MIN	**M**inimum
MLP	**M**ulti **L**ayer **P**erceptron
NB	**N**aive **B**ayes
OPT	**O**ne **T**ime **P**assword
OS	**O**perating **S**ystem
PIN	**P**ersonel **I**dentification **N**umber
RMS	**R**oot **M**ean **S**quare
RNGE	**R**ange
STD	**S**tandard **D**eviation
SUM	**S**ummary
SVM	**S**upport **V**ector **M**achines
VAR	**V**ariance
WEKA	**W**aikato **E**nvironment for **K**nowledge **A**nalysis

Chapter 1

Introduction

Since IDs can be forged easily, more complicated and sophisticated methods for identification and/or verification are needed to be used. Some human body parts are unique to only their owners such as fingerprints. So measuring this kind of body types can be used identification and/or verification purposes. To achieve this, biometrics were begun to be used. DNA matching, Ear, Eyes-Iris Recognition, Eyes-Retina Recognition, Face Recognition, Finger Geometry Recognition, Fingerprint Recognition, Odour, Typing Recognition, Signature Recognition, Vein Recognition, Gait, Palmprint, Voice Recognition etc. are only couple biometrics methods to measure the characteristics of individuals. These methods can be divided into of biometrics like DNA, fingerprints so on and behavior types of biometrics like gait, voice, so on.

Authentication factors classically fall into three categories as knowledge (what we know), possession (what we have) and inherence factors (who we are) which are called multifactor or three-factor authentication (3FA). 3FA works with something you know (ordinarily a username/password/PIN), something you have (a reliable device identifier that is not easy to copy such as OTP tokens, key fobs), something you are (a unique biometrics like retina, fingerprint) [1]. Stealing or faking all those three factors are very unlikely which makes identification and/or verification very sophisticated [2].

When mobile phones first came to our life it was used just to call or text to someone. Later on, couple features were added to cellphones like alarm and calendar. Frankly, these features weren't enough and more were implemented soon. Finally, there were too many things which can't be handled by first generation mobile phones. So as a result, smartphones were born. Today's smartphones are multifunctioning and most of the features are used for different purposes. Based on some research results, smartphone usage increases incredibly every day. In many cases, people have not just one but multiple smartphones. It's not just smartphones, tablets' usage also escalates astronomically day by day. It seems that they are like a keystone in our daily life and it's almost impossible to live without them.

This much dependency on smartphones and tablets brings security concerns inevitably. Indeed, since these devices keep so much information such as health, financial situation, business projects, personal calendar and much more, some questions marks should rise about its security. If this data switch hands from the owner to malicious attacker, results can be unpredictable. Draining checking account and savings, obtaining government documents, committing a crime or getting into legal trouble are only of a couple of things which can be done with stolen data. For this purpose, security experts desperately search better and safer security systems for smartphones and tablets.

The expression "biometrics" is gotten from the Greek words "bio" (life) and "measurements" (to measure) [3]. Even though measuring a body sounds that it's a very simple thing to do, it isn't that easy to calculate for computing systems. That was the main reason for the security engineers to come up the idea of using biometrics to ensure smartphones security. Advancements in technology made this handicap less important. In fact, using biometrics can go back to very old ages if we consider recognizing someone their faces or their unique body parts like hair or eyes. Even today, when we are not familiar with their faces we categorized that person as a stranger and strangers shouldn't be trusted. Biometrics technology for computing systems basically works form on the same concept. But unlike humans, computers need more specific information then familiarity to understand the differences. For computers, those body parts have to be unique to their owners so it can be classified by them as different recognize and separate them from each other as a trusted or untrusted user.

Smartphone biometrics is not a new topic in biometric technology. Especially accelerometers and gyroscopes sensors are the most common applied sensors for activity recognition.

While utilization of portable biometric arrangements has developed in a venture with the bigger biometrics market for quite a while, the developing universality of smartphones and the quick and sensational enhancements in their components and execution are quickening the pattern. Therefore, versatility is turning into an undeniably imperative part of the biometrics scene, and the time is on the whole correct to examine portable biometrics, and research in more noteworthy profundity how they can be utilized to their potential. To push forward the examination and application around there, an exhaustive assessment will be performed on the testing crucial and also exceptionally commonsense issues raised by the biometrics on a cellular telephone. In any case, the outcomes show that movement based biometrics is a promising venue for further study.

The rest of the thesis is organized as follows. In Chapter 2, we provide some background information on smartphones, sensors, and concepts in order to understand the study. In Chapter 3, we examine previous works on different methodologies proposed in the literature which using accelerometer or/and gyroscope information, especially those focusing on using the smartphone. In addition, we give detailed information about the number of subjects in previous studies. In Chapter 4, we provide information about our study. Therefore, we explain our applications for data acquisition. In Chapter 5, we present the conclusion of our thesis and directions for future research.

Chapter 2

Background

In this chapter, we give background information related to our study which are smartphones, sensors used in this thesis and the steps of concepts. Firstly we will describe why smartphones were used. Secondly, how accelerometers and gyroscope work. Then, the overall process will be explained.

2.1 Smart Phone Biometrics

Smartphones are getting more important in people's daily life with their better sensors and features. Since their arrival to our life, smartphones have being used more than a phone such as for gaming, emailing, navigating, listening music and much more. Using the proper applications these phones even can be used as a remote control, translator, navigator, data keeper etc. This multifunctionality can make us more productive and it is also possible that some question marks are raised in our minds about its security since these smartphones keep so many personal data in their memory. Therefore authentication is one of the main concerns for smart phone's usage.

For many years, putting a password was thought secure enough for user authentications. Even though using password/pin authentication is relatively (depends on the complexity of the password) simple and no need for extra configuration, the strength of the password determines the strength of the security and there isn't any solid identity check [4].

Recently, biometrics also involved authenticating the users since hardware industry is able to provide smaller sensors with better functionality and lower cost. Smartphones have several powerful sensors such as Global Positioning System (GPS), microphones (audio sensors), camera (image sensors), compass (direction sensors), temperature sensors together with accelerometers and gyroscopes. Apple introduced fingerprint scanning technology which was built in iPhone 5s. Samsung announced Galaxy S III smartphone with the face, voice recognition. Identify a person from their movement such as walking, jogging and running has a long history but it is the newest technology for smartphones.

Smartphones were used for this research for many logical reasons. First of all carrying a smartphone much easier and more comfortable for test subjects since all of them already have at least one instead of attaching sensors to a different part of their bodies. Another reason for using a smartphone is that their availability is higher rather than the individual sensors. Since all necessary sensors for all biometric tests in this research have integrated the smartphones it makes researcher's job simpler [5–7].

Using smartphones are very required for these days as it provides many amenities to people. For example, if they are going to somewhere and they don't know the place, smartphone's GPRS will help to them for their destination. There are many advantages of smartphones. So, smartphones are becoming more popular all over the world [7]. Smartphone usage around the world also supports our argument for using them as a test object for this research. Based on estimates of leading research and analysis firms that the world have being 2 billion smartphone users by 2016 [7]. In two years this number is expected to be two and a half billion. That means 25% increase in a short term. Figure 2.1 shows the number of smartphone users in the word (from 41 countries for making this estimation)[7].Based on the research about smartphone users by countries between 2013-2018 with 624 million users China is top of the list and it has the most smartphone users in the world[8]. That shouldn't be surprising. Even though China has the most smartphone users in the world if we compare users to country population, United States would be the first in the list. We can say smartphone usage is more common in the US than China. Since 2013, however, smartphone users increased 43.12% in China. This percentage is 38.19% from 2013 to 2016 in the US where is the second country that has the most smartphone users globally. On the other hand, regardless of their economic or politic status, all the countries in the world has a huge amount of smartphone users and this number is getting higher each day than expected. Common and highly increasing usage of smartphones besides other advantages makes it the perfect tool for our study.

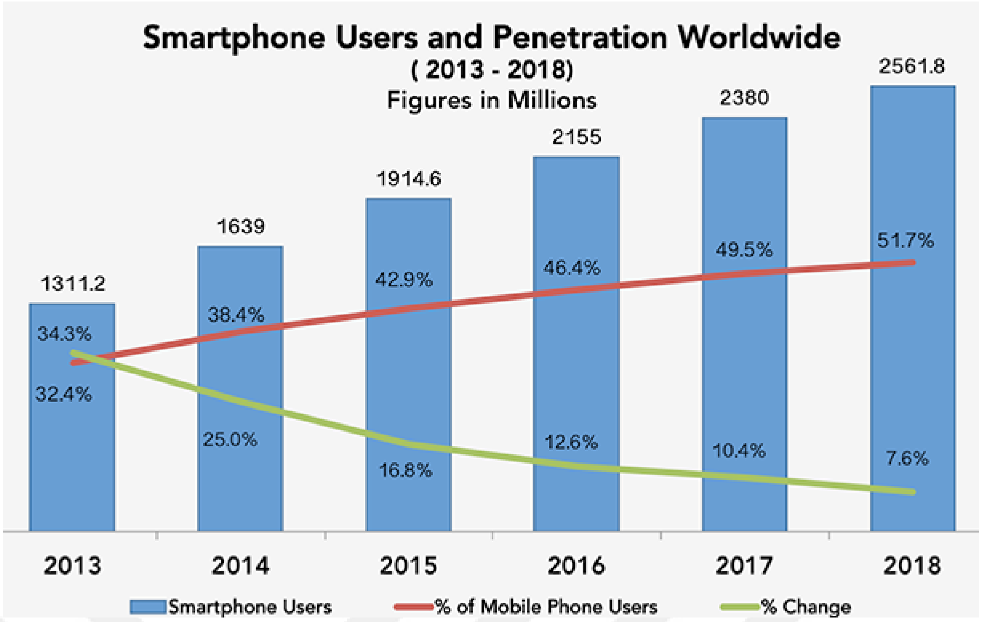

FIGURE 2.1: Penetration on SmartPhone

Following table shows top 25 countries, ranked by smartphone users between 2013 and 2018. Numbers are just predictions for 2017 and 2018 based on previous statistics. We can see that from relatively poor countries like Nigeria to most developed countries like the United States, smartphones are everywhere and its users are incredibly fast increasing. This table can gives us a clue about smartphone usage trending in the world. It also can help us for some estimations for the future.

Although smartphones are decided to be perfect tool for this research it also very important to choose right smartphone brand and model for more effective and better results. Their capacity, display, dimensions, weight, chip, platform, OS and most importantly sensors are needed to be compared to make such decision. For this purpose below compare table's research was prepared and conducted after analyzing and studying statistics and resources about most popular smartphone's features [5, 6]. They are very popular phones as they give many advantages while using it. We select to using smartphone for in the study according to a Compare Table 2.1.

TABLE 2.1: Compare Table for SmartPhones

SmartPhones Models	Capacity	Display Dimensons Weight	Chip	Platform	Operating System	Sensors
Apple iPhone 6	16/64GB/128 GB 1GB RAM DDR3	4.7-inch(diagonal) 5.44 x 2.64 x 0.27 in 129 g (4.55 oz)	A8 chip with 64-bit M8 motion coprocessor	CPU:Dual-core 1.4 GHz Typhoon (ARM v8-based) GPU:PowerVR GX6450 (quad-core graphics)	iOS 8 upgradable to iOS 10.2	Touch ID Gyro Accelerometer Proximity Light sensor Barometer Compass
Samsung Galaxy S6	16/64GB/128 GB 3GB RAM	5.1 inch 5.65 x 2.78 x 0.27 in 138 g (4.87 oz)	Exynos 7420 Octa	CPU:Quad-core 1.5 GHz Cortex-A53,Quad-core 2.1.0 GHz Cortex-A57 GPU:Mali-T760MP8	Android OS, v5.0.2 upgradable to v6.0.1	Fingerprint Gyro Accelerometer Proximity Heart rate Barometer Compass SpO2
LG Nexus 5	16/32 GB 2GB RAM	4.95 inch 5.43 x 2.72 x 0.34 in 130 g (6.28 oz)	Qualcomm MSM8974 Snapdragon 800	CPU:Quad-core 2.3 GHz Krait 400 GPU:Adreno 430	Android OS, v5.0 upgradable to v6.0	Gyro Accelerometer Proximity Barometer Compass
BlackBerry Priv	32 GB 3GB RAM	5.4 inch 5.79 x 3.04 x 0.37 in 192 g (6.77 oz)	Qualcomm MSM8974 Snapdragon 800	CPU:Hexa-core GPU:Adreno 418	Android OS, v5.1.1 upgradable to v6.0.1	Accelerometer Altimeter Gyro ToF proximity Compass

2.2 Sensors

The commonly used sensors are accelerometer and gyroscope in the word. (see Table 2.1) Most of the smartphones include accelerometer and gyroscope sensor. We can use a sensor's raw data to notice activity, gender or age recognition, and biometric identification. The main function of the accelerometer is the rate of change of celerity with time. The accelerometer has three linear accelerometers are aligned in x,y,z-axis. The main function of gyroscopes is the rate of rotation around a particular axis. Gyroscopes have tri-axial as well, making 6-degrees-of-freedom with 3-axis accelerometers. The accelerometer data is described relatively to the acceleration of gravity. The gyroscope data is described in radians per second [9].

Summarize:

- Accelerometer for changes in velocity and position,
- Gyroscope for changes in rotational velocity,

XYZ coordinates are show that the device as follows:

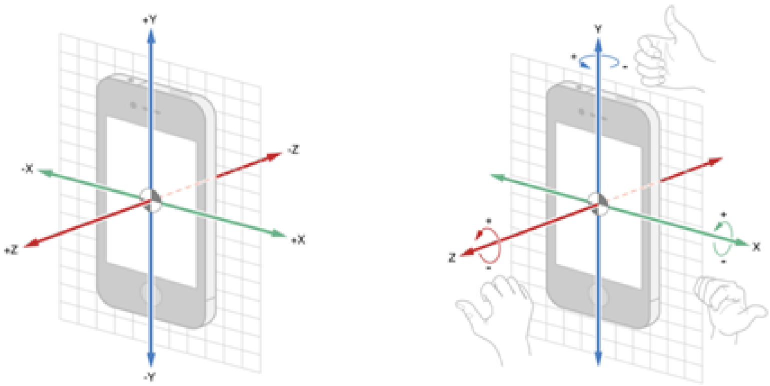

Figure 2.2: Accelerometer and Gyroscope Sensor Coordinate System

2.3 Concepts of Study

In this section, recognition process steps were expressed. These steps are; preprocessing, segmentation, feature extraction, dimensionality reduction, classification. We followed these steps for our study. Brief explanation for the steps were given below [10].

2.3.1 Preprocessing

Data that is collected from the sensors aren't just plain and contains some noise. Therefore, natural data is not suitable for analyses. Preprocessing should be applied to collected data before using it in feature extraction. The preprocessing step includes the removing of the noise and representation of raw data. Different type of methods can be used to obtain better and smoother signals for this operation such as Laplacian, Gauss, non-linear, low pass and high pass filters.

2.3.2 Segmentation

Segmentation is the process of dividing the useful information from raw data since it's really hard to get helpful information from continuous stream data. So, the signal will be split into meaningful pieces to enhance signal behavior. For this purpose, different segmentation techniques can be categorized into one of the three categories which are sliding windows, top-down and bottom-up on raw data. Sliding Windows is the most applied approach within segmentation techniques. It is uncomplicated and slight. In this technique, streaming data is separated into equal and diverse sized parts. The main difficulty with this technique is deciding the convenient size for every block. Top-down is an another segmentation technique. This method recursively divides the time arrangement until some stopping criteria are met. The bottom-up technique, on the other hand, beginning from the finest conceivable estimate merges the segments until some stopping criteria are met.

2.3.3 Feature Extraction

A big data is extracted to get a smooth vector to process much simpler. These value sets are names as feature vectors. Features can be classified by time, frequency, time-frequency and heuristic. Feature extraction technique can by-pass problems such as large amount of memory and computing power which are need to analyze massive data with many variables.

2.3.4 Classification

Classification is the final and the most significant step of the recognition process. The classification is the issue of recognizing to which of an arrangement of classifications another perception has a place, on the premise of a preparation set of information containing perceptions whose class participation is known Even though there are many, the most well-known classification methods are K-Nearest Neighbors (KNN), Decision Tree (DT), Support Vector Machines (SVM), Naive Bayes (NB), Hidden Markov Model (HMM) and Gaussian Mixture Models [11].

Chapter 3

Related Work

In this chapter, we review previously proposed methods which use accelerometer or/and gyroscope information to detect human itself, human activities, stress, fall etc. Since it is getting more popular, many researches can be found about getting accelerometer or/and gyroscope information using smartphone sensors. Common usage of smartphones with much better sensors and technology draw researchers' attentions to this area of study. While creating great input for our study, all those studies enlighten our research as well.

3.1 Previous Work

Randell and Muller added activity sensor to their GPS based Tourist Guide with X-Y accelerometer using the crossbow ADXL202 Accelerometer Evaluation Board in 2000. The data was collected from 10 people with activities standing, walking, sitting, running, walking upstairs and downstairs at a relatively low-frequency 5Hz. User's action was inferred by using a grouping algorithm and a neural system. Initial results were 85 90 % with a high precision [12].

Mantyjarvi et al. used multiple acceleration sensors with the best classification results 83-90% by training three multilayer perception neural networks using back propagation (MLP) in 2001. For this purpose, they tested the use of PCA and ICA in feature generation process with wavelet transform [13] .

In 2004, Bao and Intille gather acceleration data for their work from 20 people (13 males and 7 females) who run in age from 17 to 48 (mean 21.8, sd 6.59) with the accuracy rate of 84%. It was the first study, to examine the performance of recognition algorithms using many different accelerometers without wire. ADXL210E accelerometers were for collecting data from analog devices. Components were figured on 512 example windows of acceleration information with 256 tests covering between back to back each window represents 6.7 seconds and a sampling frequency of 76.25 Hz. C4.5 decision tree and naive bayes classifiers in The Weka Machine Learning Algorithms Toolkit were used for the study [14].

In 2005, Ravi et al. reported on their aims to user activity recognition from accelerometer data. For this purpose, they collected data with a sensor. To collect data, they selected 8 activities which are standing, walking, running, sit-ups, vacuuming, brushing teeth climbing upstairs, and downstairs. Sample frequency was 50Hz. Data were collected by the sensor which was triaxial accelerometer CDXL04M3. Subject's data handed in to an HP iPAQ wirelessly by using Bluetooth from the sensor. The data were transformed to ASCII format over a Python script. Raw accelerometer data was divided into window which sizes 256.50% overlapped on each window. Features were extracted from the each window. The extracted features were mean, standard deviation, energy, and correlation. They used WEKA toolkit for classification. Classifiers are Decision Tables, Decision Trees (C4.5), K-nearest neighbors, SVM, Naive Bayes, Stacking with MDTs, Stacking with ODTs, Plurality Voting, Bagging, and Boosting [15].

VTT Electronics workers proposed using identify people from accelerometer data in 2005. They used Analog Devices ADXL202JQ as a sensor for collecting data. The position of the sensor was behind the people's back. Their test subjects were 36 people (19 males and 17 females). The signal frequency was 256Hz. As a result, they got FRR=%5.4 and FAR= %6.4 [16].

Fitzgerald Nowlan published a paper (in 2009) about how human identification via gait recognition. For this purpose, he uses a single sensor which is composed of an accelerometer and gyroscope and collected gait characteristic data. K-nearest neighbor, naive bayes and quadratic discriminant analysis selected for classification. In this work, the test subjects were able to be identified with 95% accuracy [17].

The idea of the implementation of a real-time classification system for some basic human movements using a conventional mobile phone equipped with an accelerometer and without server processing data was presented by Brezmes et al. in 2009. Nokia N95 cell phone was used for the prototype and Python API was used to get the accelerometer's data. This study based on activities such as walking, climbing-down stairs, climbing-upstairs, sitting down, standing up, falling. Albeit most studies on subject activities' acknowledgment utilize a few accelerometers situated at specific body destinations and with particular introductions, in this study the cell phone is held by any user with no predefined introduction [18].

Spranger and Zazula make gait identification using cumulants of an accelerometer with the cell phone in 2009. Test set includes six males whose average ages were 30.2 years and average height 179. The experiment was performed on a 50 m long corridor with the surface made of stone plates. Nokia N95 was selected as a smartphone which was placed on the person's hip. Sampling data frequency is 37Hz. Feature extraction from detected gait cycles using cumulants. 1641 feature vectors generated by all cumulant coefficients from zero-lag cumulant to cumulant with lag 10. The classification was provided by support vector machines using WEKA. The average success rate was 93.1% [19].

In 2010, Kwapisz et al. published a paper about how a smartphone can be used to for person identification and authentication. They used WISDM which is a cell phone platform based on android to collect data. For their work, they acquisitioned data from 36 users who performed four activities such as walking, jogging, climb up and down stairs. Users carried smartphones in their front pants' pocket. Different models of Android phones were used such as HTC Hero, Nexus One, and Motorola Backflip for this experiment. Even though example duration was tested for 10 seconds and 20 seconds but only 10 seconds data was selected since it's more reliable. Sampling data frequency level was 20Hz. Features were generated from 600 raw accelerometer data. 43 feature vectors were generated from variations of six based feature vectors that are average, standard deviation, average absolute difference, Average Resultant Acceleration, Time between Peaks and Binned Distribution. Two classification techniques were used on WEKA which is J48 and neural networks for data mining [20].

Rasekh et al. published their paper in 2011 about human activity like walking, limping, jogging, going upstairs, and downstairs recognition system based on a 3-dimensional smartphone accelerometer. As a smartphone, they used HTC Evo. Activities were classified, tested and trained using 4 different passive learning methods as the quadratic classifier, k-nearest neighbor algorithm, support vector machine and artificial neural networks.

The maximum sampling frequency of accelerometer was 50Hz and +/- 3g sensitivity. 31 features in both time and frequency domain were generated by the accelerometer with the system gathered time series signals. These time domains are variance, mean, median, 25% percentile, 75% percentile, Correlation between Each Axis, Average Resultant Acceleration (1 resultant failure) and frequency domain are energy, entropy, centroid frequency, peak frequency [21].

In 2011, Weiss and Lockhart published the paper which is "Identifying User Traits by Mining Smart Phone Accelerometer Data". They used smartphones to predict user trait with accelerometer data. Their traits consisted of sex, height, and weight. Data was collected by their WISDM Sensor application. Example duration was 10 second and frequency was 50 ms. Dataset involved 66 subjects for gender prediction, 61 subjects for height prediction, 63 subjects for weight prediction. Feature vector was extracted average, standard deviation, average absolute difference, average resultant acceleration, the time between peaks and binned distribution. They used WEKA and classification methods were Instance Based (IB3), Neural Network (NN) and Decision Tree (J48). The accuracy of sex prediction was 71.2% at IB3 classification method, an accuracy of weight prediction was 78.9% at IB3 classification method, the accuracy of height prediction was 85.7% at NN classification method [22].

In 2014, Ferrer and Ruiz compared different algorithms for the acknowledgment of transportation modes in view of elements removed from the accelerometer information. Android application called PEATON that is able to gather GPS readings each 10-12 seconds. Frequency of accelerometer data at 1Hz. They used following smartphone models for their tests; Sony Xperia U, SonyXperia ArcS, Samsung Galaxy S, Samsung Galaxy S II and Google Nexus S. Information was gathered by 7 people over a time of three months: 3 female and 4 male. Subjects' age between 25 and 38. Members were told to completely select the travel state picked while beginning an outing (walk, motorcycle, electric tramway, train, metro, car, bicycle, bus, or wait if a participant is transferring between transport states). The length of the sliding window is set as 30 seconds without covering between continuous windows. Five models; (1) k-Nearest Neighbors (KNN), (2) Decision Trees (DT), (3) Discriminant Analysis (DA), (4) Multilayer Perceptron Neural Network (NN), (5) Recurrent Neural Network (RNN) were used in Matlab for comparing the accuracy [23].

Celenli et al. used smartphones to detect activities while performing a certain action in 2014. The activities consisted of 7 basic and 1 complex actions; standing, walking, running, jumping, standing up, sitting down, ascending stairs and descending stairs, as one action, getting in and out of a car. Each activity was 30 seconds. The subject database includes 102 persons (35 of them were females and 67 of them were males) for basic actions. Subject's age average was 30. 30 subjects performed for complex action. IOS application was developed which includes accelerometer and gyroscope sensors. iPhone 5 and iPhone 5s smartphones were used in this approach so as to data acquisition. Frequency was chosen 100 Hz. Phone locations or the place of the action weren't specified. C++ code was developed for feature generation from collected data. Feature vector was extracted min, max, mean, Root Mean Square, Standard Deviation, Binned-average, Zero Crossings. They used WEKA toolkit classifiers which include regression, Bagging, Multi-layer Perceptron, K-Star, Bayesian Network, Logistic Model Tree for classification. As a result, K-star gave to recognition rates exceeding 98% [24].

In 2014 Aguiar et al. developed Android application which is ADLS for fall detection using accelerometer data. When a fall was detected, sound the alarm of application. Their test subject set includes 36 people, 28 of them are young people and 8 of them are older people. 24 males and 4 females of young people with average age of 25, the average height of 175, average weight of 71. 4 males and 4 females of older young people with average age of 66, average height of 175, average weight of 72. They tested three different classifiers of offline machine learning tool: Decision Trees, K-Nearest-Neighbors, and Naive Bayes. They obtained feature vector from mean, median, maximum, minimum, root mean square, standard deviation, median deviation, interquartile range, energy, entropy, skewness, and kurtosis. The success of their fall detection algorithm was 97.5% [25].

Jain and Kanhangad proposed, in 2015, a method using biometrics for user authentication which achieves the lowest EER of 0.31%. There were 104 participants. The dataset includes 9 users between the age 31 and 36 years, 40 users between 26 and 30 years and 55 users between 19 and 25 years. IntelliJ IDEA android application was developed and ran it on Samsung Galaxy S-II GT-I9100. Samsung Galaxy Note-II N7100 was used as well to acquire dataset that contains data from 30 subjects [26].

Osmani et al. proposed in 2015 to use smartphones for detecting behavior changes from accelerometer data. Their test has been done on 6 persons for 10 months at the psychiatric hospital in Hall in Tirol, Austria. The authors reported that they could gain success average precision was 81%, and recall was 82% by using Naive Bayes, k-nearest neighbor, j48 search tree, and a conjunctive rule learner. They evaluated the test with the Gaussian distribution method and increased the success rate average precision to 96%, and recall to 94% [27].

In 2015, Garcia-Ceja et al. achieved a maximum overall success rate of 71% for user-specific models and a success rate of 60% for the use of similar users models with 30 subjects (18 [60%] males and 12 [40%] females) for automatic stress detection in working environments from smartphones' (using the built-in sensors of Samsung Galaxy SIII Mini smartphones) accelerometer data. In order to extend the battery life, they set the accelerometer sampling rate at 5 Hz. They used 4 classifiers: Naive Bayes; Decision Tree; Ordinal Naive Bayes for their experiments. Accelerometer information was used to portray subjects' conduct by extricating time domain and frequency domain [28].

Another paper that was published in 2015 by San-Segundo et al. aimed to get feature extraction from smartphone inaction signals for human activity segmentation. Those segmentations were six different physical activities; walking, walking-upstairs, walking-downstairs, sitting, standing and lying. Input data was ensured, 30 volunteers. This dataset has been grouped into six subsets (randomly) because performing a six-fold cross-validation procedure. The dataset which was called UCI Human Activity Recognition with Smartphones data cluster shared publicly. They stated that Activity Segmentation Error Rate as lower than 0.5%. It was recommended that defining new frequency warping strategies and focusing on evaluating these feature extraction operations in a different application: gait recognition instead of activity recognition as future work [29].

Biometric Authentication Technique Using Smartphone Sensor was published in 2016 by Laghari, Rehman, and Memon. The paper presents authenticating biometrically with the help of smartphone's motion sensor. The signal matching concept was used for identification. They concluded their experiment result as 6.87% FRR and 1.46% FAR. As they mentioned the method they used can be improved and more accurate results can be gathered from frequency analysis of the signature signal for authentication process [30].

To compare ours to previous works can help describing and explaining the differences between studies for clear and better understanding and also it may point new research ideas. To achieve this Table 3.1 was created. At this table, columns are corresponded as follows; reference to the work, sensor type, using device, proposed method, the number of subjects data, sampling rate, overall best success among different classifiers' results, feature vectors, and classifiers.

There are 4 main contributions of our work. Firstly, we collected data from a large set of subjects with varying ages and gender. Secondly, data were acquisitioned from subjects at different environment.Thirdly, the application was developed at two different platforms which are IOS and Android for getting data. Lastly, more than one recognition analysis were presented.

TABLE 3.1: Compare Table for SmartPhone

	Sensor	Device	Proposed Methods	Subjects	Frequency	Overall Best Success	Feature Vektors	Classifier
RM00 [12]	Portable Device	ADXL202	Action Recognition	10	5Hz	85-90%	See Paper	Neural Network
MHS01 [13]	Multiple Sensor	N/A	Action Recognition	N/A	256Hz	83-90%	See Paper	See Paper
BI04 [14]	Analog Device	ADXL210E	Action Recognition	20	76.25Hz	84%	Mean Energy Frequency Entropy Correlation	Decision Table Nearest Neighbor C4.5 Decision Tree Naive Bayes
RDM+05 [15]	Analog Device	Hoarder Board Bluetooth enabled HP iPAQ	Action Recognition	N/A	50Hz	See Paper	Mean Standard Dev. Energy Correlation	Decision Table C4.5 Decision Tree Naive Bayes Nearest Neighbor SVM
MLV+05 [16]	Analog Device	ADXL202JQ	Human Identification	36	256Hz	84.6%	See Paper	See Paper
N09 [17]	Analog Device	N/A	Human Identification	N/A	50Hz	95%	See Paper	K-nearest Neighbor Naive Bayes
BGC09 [18]	Mobile Phone	Nokia N95	Action Recognition	N/A	N/A	90% (for walking)	See Paper	K-nearest Neighbor
SZ09 [19]	Mobile Phone	Nokia N95	Human Identification	6	37Hz	93.1%	Cumulant Coefficients	SVM

KWM10 [20]	Smartphone (Android)	Nexus One HTC Motorola	Human Identification	36	20Hz	78.6%	Average Standard Dev. Ave.Absolut-Diff. Average Resultant Acce. Time between Peaks Binned Distribution	J48 Neural Networks
RCL11 [21]	Smartphone (Android)	HTC Evo	Action Recognition	N/A	50Hz	84.4%	Variance Mean Median Correlation Axis Average Resultant Acce.	Quadratic K-NN SVM Artificial Neural Networks
WL11 [22]	Smartphone (Android)	N/A	UserTraits Identification (gender, weight, height)	70	50Hz	85.7%	Average Standard Dev. Ave.Absolut-Diff. Average Resultant Acce. Time between Peaks Binned Distribution	Instance Based (IB3) Neural Network (NN) Decision Tree (J48)
FR14 [23]	Smartphone (Android)	Sony Xperia U SonyXperia ArcS SamsungGalaxy S Google Nexus S	Travel Behavior Recognition	7	1Hz		Mean Standard Deviation Max. Acceleration Min.Acceleration First and Third Quartile	K-NN Decision Trees Discr.Analysis Neural Network Recurrent NN

CSE+14 [24]	Smartphone (IOS app)	iPhone5 iPhone 5S	Action Recognition	102	100Hz	98%	Min, Max, Mean, Standard Dev. Root Mean Square Zero Crossings Binned-average	Regression Bagging Multi-layer Per. K-Star Bayesian Network Logistic M. Tree
ARS+14 [25]	Smartphone (Android)	N/A	Fall Detection	36	4Hz	97.5%	Mean,Median Max, Min. Root Mean Square Standard Dev. Median Dev. Quartile Range Energy, Entropy Skewness,Kurtosis	Decision Trees K-NN Naive Bayes
JK15 [26]	Smartphone (Android)	SamsungGalaxy S-II GT-I9100 Note-II N7100	Human Identification	104	N/A	98.5%	Point Curvature Swipe Curvature	See Paper
OGB+15 [27]	Smartphone (Android)	N/A	Stress Detection	6	N/A	81%	See Paper	Naive Bayes K-NN J48 Conjunctive Rule
GOM15 [28]	Smartphone (Android)	Samsung Galaxy SIII Mini	Stress Detection	30	5Hz	71%	See Paper	Naive Bayes Decision Tree Ordinal NB

3.2 Database

In this section, we summarize the subjects of related studies. Thus, we specify the number of subjects and their properties. In the following Table 3.2 presented the number of subjects in previous work and our work. Columns in the table correspond to the following: reference to the work, a number of the subject's data was collected from, gender and average age, respectively.

Table 3.2: Related Works' Database

	Subjects	Male	Female	Average Age
RM00 [12]	10	N/A	N/A	
BI04 [14]	20	13	7	21.8 (Between 17 and 48)
MLV+05 [16]	36	19	17	N/A
SZ09 [19]	6	6	0	30.2
KWM10 [20]	36	N/A	N/A	N/A
WL11 [22]	66	38	28	Between 18 and 24
FR14 [23]	7	4	3	Between 25 and 38
CSE+14 [24]	102	67	35	30 (Between 19 and 55)
ARS+14 [25]	36	24	4	28+(-)3
JK15 [26]	104	N/A	N/A	Between 19 and 36
OGB+15 [27]	6	N/A	N/A	N/A
GOM15 [28]	30	18	12	37.4+(-)7.26
Y09 [31]	4	N/A	N/A	N/A
SLZ+10 [32]	7	6	1	Between 25 and 46
KMM10 [33]	29	N/A	N/A	N/A
HTM11 [34]	5			
DDK+12 [35]	10	N/A	N/A	N/A
UIE13 [36]	20	16	4	Between 18 and 59

Chapter 4

Recognition Process

The main purpose of our study is detecting the gender, age and identification of a person with high certainty, using the accelerometers on a mobile device such as cell phone and/or tablet. Steps for gender, age and identification detections are explained in this chapter. The details of accelerometer usage on the Android and IOS platform are clarified. We also explain how detection process is implemented.

Smartphones are very powerful devices with their sensors that are built into them. Sensor structure gives us a chance to get to many sorts of sensors. Our study bases on these sensors. On the other hand, while we do our research on mobile devices, one of the biggest problems we had to face was power limitations of for these portable gadgets mobile devices is one of the most important boundaries. Battery drain can makes the experiments more harder than it's. These gears can consume even more electricity when the sensor are used. Some sensor can exhaust much more power than others. These power consuming differences are significantly huge sometimes. For this purpose, iInstead of using power consuming sensors like GPS, we used only the accelerometer to defeat this complication. Accelerometers are the sensors which allow us to detect age, gender, and identity. Less power absorption isn't the only advantage working with accelerometers but high accuracy as well.

As mentioned before the main focus of this thesis is to be able to recognize age, gender, and identity. Three different positions; right, left and back pockets are chosen for smartphones and one position; hand-held is chosen for the tablet to collect data. After we collected the data from various positions to frame the training model, preprocessing segmentation, feature extraction and classification steps are followed. Walking and running are the chosen to study on the activities.

Two application were developed to gather the data from test subjects. One is for Android platform and another one is for IOS platform. More details about these applications are stated in this chapter. After the development of these apps, there were installed to mobile devices and from different age and gender groups in different locations, data is collected. Minutes of different subjects are explained in the following subtitles as well. To sort out the collected data, we used MATLAB. Feature vectors were tested in different classes.

4.1 Applications

To monitor the motion of an android device there are couple sensors that are provided by the Android platform. While the accelerometer and gyroscope are hardware-based sensors, the gravity, linear acceleration, and rotation vector sensors can be either hardware or software based [37]. Since to extract their data, the software-based sensors rely on one or more hardware-based sensors the availability of them is more variable. Motions sensors (gravity sensors, accelerometers, gyroscopes and rational vector sensors), environmental sensors (barometers, photometers, and thermometers) and position sensors (orientation sensors and magnetometers) are three supported broad categories of sensors [38].

3-axis (x, y, z) coordinate system is used to express data values by the sensor framework. X axis is horizontal and points to the right, the Y axis is vertical and points up, and the Z axis points toward the outside of the screen face when its default orientation [39].

Following sensors use this coordinate system;

- Acceleration sensor,
- Gravity sensor,
- Gyroscope,
- Linear acceleration sensor,

- Geomagnetic field sensor

The most critical point to comprehend on this coordinate framework is that the axes are'nt swapped when the gadget's screen introduction changes-that is, the sensor's coordinate framework never shows signs of change as the gadget moves. This conduct is the same as the conduct of the OpenGL coordinate framework [39, 40].

Another point to comprehend is that your application mustn't expect that a gadget's common (natural) orientation is portrait. The common orientation for some tablet gadgets is scene. What's more, the sensor coordinate framework is constantly taking into account the regular orientation of a gadget [37, 39].

At last, if your application links to sensor information to the on-screen show, you have to utilize the getRotation( ) method to decide screen rotation, and afterward use the remapCoordinateSystem( ) technique to guide sensor coordinates to screen coordinates. You have to do this regardless of the possibility that your show determines portrait-only display [37, 38].

Sensor: TYPE_ACCELEROMETER

Sensor Event Data;

- SensorEvent.values[0]: Acceleration force along the x axis (including gravity).
- SensorEvent.values[1]: Acceleration force along the y axis (including gravity).
- SensorEvent.values[2]: Acceleration force along the z axis (including gravity).

Sensors and get raw sensors are accessible by using data the Android sensor framework. The sensor framework is in the android.hardware package. It includes some classes and interfaces [41, 42]. These are described in below:

SensorManager

You can create an example of the sensor service with SensorManager class [38]. The following operations can be performed with this class :

- Access sensor

- Listing sensor
- Registering and unregistering sensor event listener
- Obtain orientation knowledge
- Calibrate sensors
- Reporting sensors precision
- Determine data collection rates

Sensor

You can create an example of the specific sensor with Sensor class. You can determine what a sensor can do with this class. [38].

SensorEvent

The system uses this class to create a sensor event object, which provides information about a sensor event [38]. The parts of the sensor event object are the following information:

- The coarse sensor data
- The accuracy of the data
- The timestamp for the event
- The type of sensor which generated the event

SensorEventListener

This is a interface which you can create sensor events. You can use two callback methods that receive notifications: sensor values change, sensor accuracy changes [38].

When we take a look at iOS application [43]; iOS provides a CMMotionManager object as the gateway for motion services [37]. The following operations can be provided with this services :

- Application with accelerometer data

- Application with rotation rate data
- Application with other device-motion data
- Application with the type of sensor which generated the event

Subsequent to making an occurrence of CMMotionManager, an application can use it to get four sorts of movement [44] :

- Raw accelerometer data
- Raw gyroscope data
- Raw magnetometer data
- Processed device-motion data

To handle movement information by occasional testing, the application calls a "start" strategy taking no contentions and intermittently gets to the movement data held by a property for a given sort of movement data. This methodology is the prescribed methodology for applications, for example, recreations. Taking care of accelerometer information in square presents extra overhead, and most amusement applications are intrigued just the most recent example of movement data when they render a frame.

Accelerometer. Call startAccelerometerUpdates to begin updates and periodically access CMAccelerometerData objects by reading the accelerometerData property.

In this study, data was acquisitioned from accelerometer sensors via two mobile application (see Figure 3). Data collection was controlled by our iOS application ([43]) for details about deriving sensors' data on iOS as explained above) which the work described in our paper [24] and Android application was developed (see [38] for details about deriving sensors' data on android as explained above) as free, open source, progressing quite rapidly and phone options are diverse. This project employes iOS phones; including iphone 5, iphone 5S, iphone 6 and Android phone; including LG Nexus 5 for collecting data.

Applications have some general features:

- Frequency levels 100 Hz,

- Each for a period of around 30 seconds,
- Finish sound for end of period,
- The application delays retrieving data for 5 seconds for eliminate noise occurrences at the begining of the activity,
- Collecting data is saved as a .txt file,
- Txt file is saved smartphone's sd card for Android,
- Txt file is sent to mail for IOS,
- File name; person id-action name
  - Android: 1.1_walking.txt
  - iOS: 1.1-walking_acc.txt
- For example accelerometer file for person 1 (as it seen below):

FIGURE 4.1: File Name

- Txt files (for both iOS and Android) contains 7 columns that First column is x axis, second column is y axis, third column is z axis, fourth column is timestamp, fifth column person name, sixth column age group,and finally seventh column represents gender (Figure 4.2 and Figure 4.3) .
- Interface for iOS has 4 blank fields that need to be filled. Those fields are person ID, name, age, and gender. There is also a scroll down menu where one of two activities (walking, running) should be chosen (even though, there are more activities at scroll down menu only walking and running were used for this study). At the bottom of the interface, there is start button to run the program and post button to send the results to the email. Interface for Android has 4 blank fields that need to be filled as well. Those fields are same with iOS interface (person ID, name, age, and gender). Two buttons to start walking (walking start) activity and to start running (running start) are located at the top of the interface. All data is restored to sd card at Android devices (Figure 4.4).

```
-0.05 0.55 0.23 35827.269294 TunahanSevis young male
-0.11 0.59 0.21 35827.279172 TunahanSevis young male
-0.19 0.63 0.17 35827.289052 TunahanSevis young male
-0.29 0.62 0.15 35827.298931 TunahanSevis young male
-0.37 0.58 0.11 35827.308810 TunahanSevis young male
-0.43 0.51 0.05 35827.318689 TunahanSevis young male
-0.42 0.41 0.06 35827.328568 TunahanSevis young male
-0.30 0.28 0.07 35827.338436 TunahanSevis young male
-0.11 0.17 0.06 35827.348325 TunahanSevis young male
0.10 0.16 -0.05 35827.358183 TunahanSevis young male
0.37 0.25 -0.19 35827.368071 TunahanSevis young male
0.65 0.50 -0.44 35827.377961 TunahanSevis young male
1.00 0.97 -1.01 35827.387837 TunahanSevis young male
1.44 1.65 -2.20 35827.397703 TunahanSevis young male
1.77 2.33 -2.81 35827.407570 TunahanSevis young male
1.78 2.47 -2.05 35827.417471 TunahanSevis young male
1.53 2.11 -1.02 35827.427336 TunahanSevis young male
1.14 1.60 -0.49 35827.437229 TunahanSevis young male
0.85 1.06 -0.36 35827.447081 TunahanSevis young male
0.77 0.58 -0.46 35827.456969 TunahanSevis young male
0.70 0.20 -0.53 35827.466856 TunahanSevis young male
```

FIGURE 4.2: IOS .txt File

```
1.21	-14.66	-0.7	357565614878992	TunahanSevis	young	male
1.32	-14.25	-4.7	357565624949793	TunahanSevis	young	male
-0.93	-14.32	-7.34	357565635020594	TunahanSevis	young	male
-1.22	-13.52	-5.9	357565645091395	TunahanSevis	young	male
-0.46	-11.98	-8.48	357565655162196	TunahanSevis	young	male
-2.43	-10.69	-2.34	357565665232996	TunahanSevis	young	male
-5.96	-10.29	6.65	357565675303797	TunahanSevis	young	male
-7.52	-10.0	4.58	357565685374598	TunahanSevis	young	male
-7.51	-10.47	-0.65	357565695445399	TunahanSevis	young	male
-6.08	-10.23	-4.94	357565705516199	TunahanSevis	young	male
-4.69	-9.39	-5.79	357565715587000	TunahanSevis	young	male
-2.71	-8.48	-2.76	357565725657801	TunahanSevis	young	male
-0.55	-7.79	-1.09	357565735728602	TunahanSevis	young	male
1.71	-7.37	-1.46	357565745799403	TunahanSevis	young	male
3.22	-7.17	-1.88	357565755870203	TunahanSevis	young	male
4.15	-6.91	-1.55	357565765941004	TunahanSevis	young	male
4.07	-6.5	-1.16	357565776011805	TunahanSevis	young	male
3.04	-6.17	-0.45	357565786082606	TunahanSevis	young	male
2.57	-5.79	0.17	357565796153406	TunahanSevis	young	male
```

FIGURE 4.3: Android .txt File

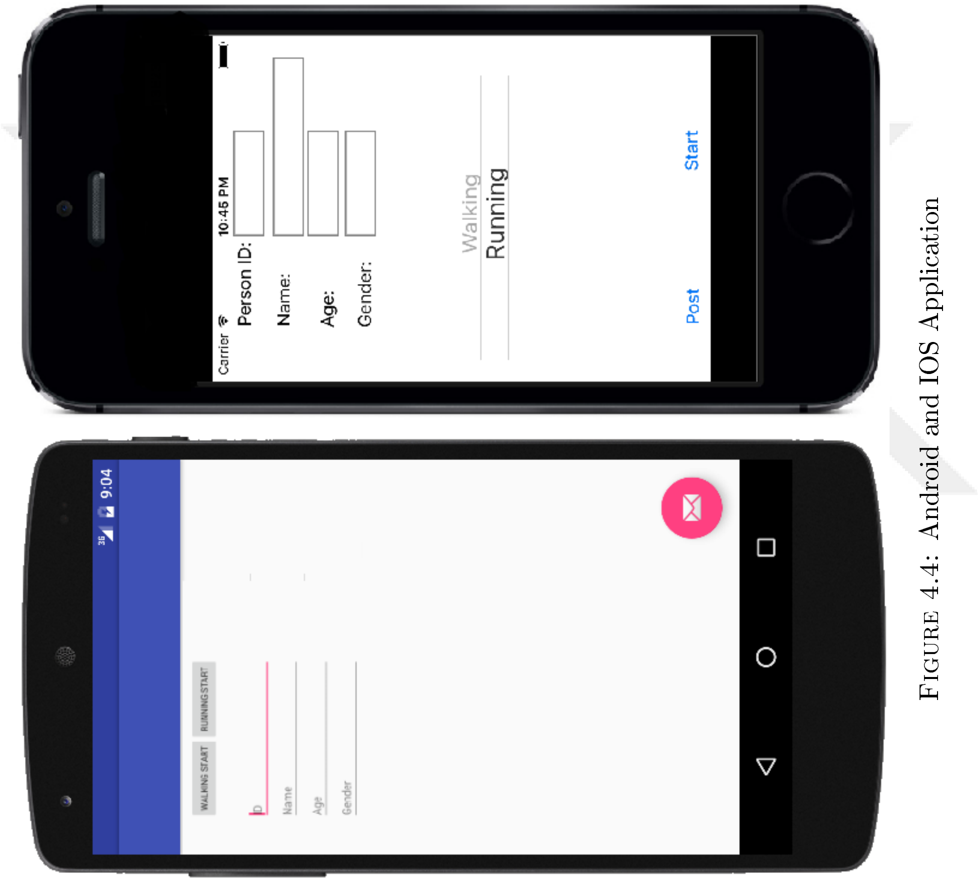

Figure 4.4: Android and IOS Application

4.2 Data Collection

As stated previously, the objective of our thesis is to achieve to detect age, gender, and identification of people with high accuracy using portable devices. To be able to gather this kind of data from such devices, two applications were developed and used. One of this application is installed to IOS devices and the other one is installed to Android devices. The very first venture of our study is the gathering of accelerometer information from different subjects while they are playing out the given exercises.

A variety of different age and gender groups, operating systems (IOS and Android), and activities were chosen specifically to get more challenging result for better results. Preferring different locations for portable devices during activities were taking place by the subjects is another reason to bring even more challenge to our study.

We collected data activity from smartphones for both IOS and Android base for age, gender and identification detection. In this study, we use smartphones which are LG Nexus 5, Iphone 5/5S/6 with a built-in triaxial accelerometer and we use a tablet which Samsung Galaxy Note 10.1. We use the tablet for a small dataset collection because we want to analyze how hand-held position affects the test results. 100 Hz, as the maximum sampling rate of the accelerometer, was used by applications for the data collection.

We attempted to make true situations however as could reasonably be expected while gathering information. So, we did'nt indicate any clothing, shoes or places. We did not interfere with the orientation of the phone.

For gender recognition, using IOS-based mobile phones, 58 person (25 female, 33 male) performed running activity and 90 person (35 female, 55 male) performed walking activity. For age detection in IOS systems, 90 people (54 young, 36 adult) performed walking and 60 people (39 young, 21 adult) involved the experiment for running.

Table 4.1: Number of Subjects for Identification-Iphone 5/5S/6

	Subject	Female	Male
Walking	110	42+(-)3	68+(-)3
Running	90	36+(-)3	54+(-)3

For Android based applications mobile phone, 34 person performed walking activity and 20 person performed running activity. All of them were used for age recognition, gender

Table 4.2: Number of Subject for Gender Recognition-Iphone 5/5S/6

	Female	Male
Walking	35	55
Running	25	35

Table 4.3: Number of Subject for Age Recognition-Iphone 5/5S/6

	Young	Adult
Walking	54	36
Running	39	21

recognition, and identification. Detailed information is provided for each of the following Tables 4.4, 4.5, 4.6.

Table 4.4: Number of Subjects for Identification - Nexus 5

	Subject	Age Average
Walking	34	32,6
Running	20	30,4

Table 4.5: Number of Subject for Gender Recognition-Nexus 5

	Female	Male
Walking	15	19
Running	7	13

Table 4.6: Number of Subject for Age Recognition-Nexus 5

	Young	Adult
Walking	18	16
Running	12	8

Android based tablet was used for identification recognition. Walking data was collected from 15 female and 19 male (34 subjects in total). Age average of subjects for the tablet is 32,6. Running data was collected from 7 female and 13 male (20 subjects in total). Age average of subjects for tablet is 30,4. Since the tablet can't fit most of the pockets, subjects hold the tablet during the performance of the activities.

Table 4.7: Number of Subjects for Identification-Samsung Galaxy Note 10.1

	Subject	Age Average
Walking	34	32,6
Running	20	30,4

4.3 Preprocess

Preprocessing step is about taking the selected data into a form that it can be worked with [45]. In the beginning, the data may not be in a format that is applicable to work. The more consistent and better results likely to be achieved with preprocessing. As known, high-quality data will lead to the high-quality conclusion.

Noise and outliers, missing values, and duplicated data should be determined. Irrelevant and redundant or noisy and unreliable data can make much harder to reach accurate conclusions. In this step, missing data may need to be removed or fixed and noise should be cleaned by applying different kinds of operations such as low pass, high pass, Laplacian and Gaussian filters. Smoothing the signal and converting an unequally-sampled time series to an equally-sampled one can be done by using these signals. If there is any, sensitive information may need to be anonymized or removed as well [45].

At preprocessing step, as mentioned before, there is noise with collected data. Situations such as after starting the app till the smartphone is replaced at users' pocket or when mixing activities with confusion like walking instead running or vice versa, cause noise. To eliminate the noise at the beginning, the app begins 5 seconds later after the start button was hit, for both iOS and Android. When the data is collected, the graph was drawn with the help of MATLAB to analyze other noise if there is any.

4.4 Segmentation

Data segmentation is the process of taking data and segmenting it so that the data can be used more effectively and efficiently and also it can be implemented easily. A continuous signal needs to be divided into data windows to be able to extract features by leaving no gaps between consecutive windows. That's the reason that segmentation is also named as windowing. Fixed-length sliding windows are used to achieve this aim. The main purpose of the segmentation is to make it easier to extract features from the signal. There are different techniques for segmentation such as sliding windows, top-down and bottom-up. Even though, each technique has their advantages and disadvantages, sliding windows have the most popularity in comparison to others. Because they are simple and lightweight. These are one of the reasons that we used this technique for our study. In this approach, streaming data is divided into equal or varied size pieces. Determining the optimal size and overlapping ratio of windows is critical since it has a direct effect on classification performance [46].

Since it is really hard to analyze streaming data, the data is divided into windows. To get necessary information for our study, we divided data 2-second windows. 1, 2, 4 and 6 were tested at our previous work (see at [24]) as window sizes and 2 seconds was found as the optimal windows size. Each window was overlapped with the previous one with 50 percent of it. So 2-second windows were overlapped by 1 second.

4.5 Feature Extraction

Feature Extraction creates a set of features by disintegrating the initial data. A feature is a combination of aspects that is of special interest and captures important characteristics of the data. Describing data with a far smaller number of attributes than the original set can be done by using feature extraction. A huge amount of memory and computation power or a classification algorithm can be required without extracting a set of features from a large data window. The ways to compute features are time, frequency and time-frequency analysis.

We used feature vectors that are in the recognition process is explained in the Table 4.8 as shown below [47]. MATLAB code was developed to extract a feature vector from each window. The original vector comprised of the values of minimum, maximum, mean, standard deviation, root mean square, range, kurtosis, skewness, variance, sum and resultant values from the accelerometer data. The base feature vector will be referred as feature vector of size 31.

Table 4.8: Feature Vectors

FV	Accelerometer	Explanation
min	(x,y,z)	The smallest values in acceleration data along all three axes
max	(x,y,z)	The largest values in acceleration data along all three axes
sum	(x,y,z)	Absolute value of addition
range	(x,y,z)	range(a) returns the difference between the maximum and the minimum of a sample.
mean	(x,y,z)	The average values in acceleration data along all three axes
var	(x,y,z)	The expectation of the squared deviation of a random variable from its mean
standart deviation	(x,y,z)	Dispersion of a set of acceleration data along with all three axes values.
kurtosis	(x,y,z)	Measure of the tailedness of probability distribution of a real valued random valiable
skewness	(x,y,z)	Measure of the asymmetry of probability distribution of a real valued random valiable
root mean square	(x,y,z)	The square root of average of the square values in acceleration data along all three axes
resultant	(x,y,z)	The square root of the sum of the squares of the mean of the three axes

4.6 Classification

Classification is an action that assigns items in a collection to target categories or classes. To precisely predict the target class for each case in the data is the purpose of classification. Classification task starts with an information set in which the class assignments are known. By comparing the forecasted values to known target values in a set of test data, classification models are tested.

Naive Bayes, KNN, Bagging and Decision Tree classfiers are some techniques for classification. Decision trees are a forceful classifier. KNN is known among the oldest non-parametric classification technique. Naive Bayes which is based on the principle of Maximum A Posteriori (MAP) is very outstanding algorithm. Even though there are more techniques such as HMM, GMMs or ANN aren't preferred because of their computational complexity either at training or classification steps. We use Weka Data Mining Tool for classification [48].

We have examined the classification methods used in other articles described in Chapter 3. The classification methods that we used for our study is explained below:

Bagging: Bagging stands for Bootstrap Aggregating so its know with that name as well. It is way that is used for statistical classification and regression to improve the stability and accuracy of machine learning algorithms. In this method;

- parallel ensemble: each model is built independently
- aim to decrease variance, not bias. the variance is decreased for avoiding overfitting.
- suitable for high variance low bias models (complex models)
- an example of a tree based method is random forest, which develop fully grown trees (note that RF modifies the grown procedure to reduce the correlation between trees).

The most known example of bagging method is Ozone data. To analyze relationship between ozone and temperature this method was used.

J48: J48 algorithm which is also called C4.5 is used to generate a decision tree. In reality, J48 is an open source Java implementation of the C4.5 algorithm in the Weka data mining classifier. Based on another description, J48 is a class for generating a pruned or unpruned C4.5 decision tree. This method is a very powerful decision method used for classification and performs well on the lonosphere dataset. In 2008, Springer LNCS ranked C4.5 as number 1 in the 'Top 10 Algorithms in Data Mining' paper.

K: K is one of the Lazy classifiers in Weka which are the most useful for datasets with few attributes. According to Weka, Kstar is an instance-based classifier, that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function. It differs from other instance-based learners in that uses an entropy-based distance function. In this method, new data instances are assigned to the class. Then entropic distance is used to retrieve the most similar instances from the data set. Using random over sampling techniques, Kstar model was also to achieve higher predictive sensitivities and specificities.

Multilayer Perceptron: MLP maps sets of input data onto a set of appropriate outputs.In this method, there are multiple layers of nodes and each layer fully connected to the next one. Using a learned non-linear transformation, the input is first transformed. The input data is projected into space where it becomes linearly separable by this transformation.

It is referred as a hidden layer. Sufficient MLPs a universal approximator are made by this hidden layer. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable.

NaiveBayes: NaiveBayes Bayes is a simple probabilistic classifiers classification based on Bayes' Theorem with strong independence assumptions between the features. All properties independently contribute to the probability. Naive Bayes is easy to build and particularly for very large data sets and has the best of it even highly sophisticated classification methods. To use naive bayes method, first the data need to be converted into a frequency table. Then, by finding the probabilities, likelihood table is created. Finally, for each class, Naive Bayesian equations is used to calculate the posterior. The highest posterior probability is the outcome of prediction.

Chapter 5

Experimental Analysis Results

The main purpose of this chapter is to evaluate the performance of the proposed system based on selected system. Two different kinds of activities (walking and running) are chosen to be performed for the tests in two batches. In the upcoming sections, test designs will be explained in detail. As mentioned, the data was collected by the two applications (one of them was developed for Android OS and the other one is for IOS). For each activity, we collected approximately 40-second data from each subject. After the noise was eliminated, it was reduced to around 30 seconds. Then, these collected data was given to WEKA machine learn software to analyze of age, gender, and identification of the subjects. Waikato Environment for Knowledge Analysis (WEKA) is a well known open source (licensed under the GNU General Public License) machine learning software, developed at the University of Waikato in New Zealand. WEKA has the ability to support data preprocessing, clustering, classification, regression, visualization, and feature selection. This chapter was concluded by providing analyze results done by WEKA.

5.1 Performance Metric

Application Platform: Two different types of platform (Android and IOS) were preferred to achieve a more authentic conclusion. Both operating systems have them the favorable and the unfavorable factors or reasons to be used or not. Collecting data from these operating system bring our study, even more, challenges.

Feature: Feature vectors were determined based on test results. We split feature vectors into two categories extended and basic. Feature vectors have an important role for solid results.

Mobile Device: Sensors which are embedded into mobile devices can vary based on mobile devices' brand and even model. This variety has a direct effect on the performance of sensors. The different sensor can produce different noises and data quality can change depending on the sensor quality. Plus, the computational power of a mobile device is another factor which may directly affect the performance.

Activity: We attempt to perceive the objective activity that was being performed amid tests. Chosen action sets are additionally imperative regarding precision of the framework. It is possible to affect the recognition performance by adding new activities to an ongoing system.

Window Size and Frequencies: Aside from immobile activities, there is the pattern for every individual activity. Windows size is the span amid. In this duration only without performing any classification data is collected. This kind of data is very important to identify activities. In our previous study, we tested different frequencies and windows sizes but for this thesis, we used fixed windows size and frequencies as the windows size is 2 and the frequencies 100 Mhz.

Classifiers: One of the most essential strides for activity recognition process is classification. For this reason, classifier has a vital part for more accurate results and for same reason, the most appropriate classifiers were chosen for our system.

Environment: The data was collected in different kind of environment platforms such as flat, fixed, sloping, and stony etc. Choosing various type of environmental factors has direct effects to get successful results.

Gender and Age: The data which depends on age and gender can lead us to disparate conclusions that can change the outcome. Instead of choosing only one age group or gender, we decided to prefer different types of genders (female and male) and different types of age groups (between 15 - 57) for more accurate and better results.

5.2 Experimental Analysis Results of Identification

As it was described in chapter 4, firstly we collected data with an application for IOS and Android. Then we extracted feature vectors from each sliding window segmentation. In this chapter, these features have performed the tests by using WEKA data mining toolkit for identification. We used 6 different classifiers which were applied to different feature vectors were tested. The classifiers are Naive Bayes, Decision Tree, Bagging, Multilayer Perceptron, K-star. The performance of chosen classifier was measured using test options; 10-fold cross validation and supplied test set. Default windows size for all the classifiers is 2 seconds, frequency is 100 Hz (200 sample) each window overlapped with the previous one with half the size of a window.

As mentioned in data collection chapter, with IOS based application, the walking data was collected from 110 test subjects. In Table 5.1, test result based on using 5 different classifiers and 10-fold cross validation were shown with feature vectors which were from 110 people. Test result, time and used features were also shown in this table. Similar to the previous studies, K star algorithms are the algorithm for identification systems that gives the best results with overall success rate 94.7747%. Naive Bayes algorithm that is used for most studies and which gives results very quickly, accomplished 84.407% success rate. More details which were gathered from WEKA also shown in the table below.

Table 5.1: Identification Success Rate from Walking Data (1)

<table>
<tr><th>Classifier</th><th>Success</th><th>Time Taken to Build Model</th><th>Features</th></tr>
<tr><td>Naive Bayes</td><td>84.407 %</td><td>0.05 seconds</td><td rowspan="5">min1,min2,min3
max1,max2,max3
sum1,sum2,sum3
rnge1,rnge2,rnge3
mean1,mean2,var1
mean3,var2,var3
std1, std2, std3
kurtosis1,kurtosis2
kurtosis3,skewness1
skewness2,skewness3
rms1, rms2, rms3
resultant</td></tr>
<tr><td>Decision Tree (J48)</td><td>85.4852 %</td><td>1.65 seconds</td></tr>
<tr><td>Bagging</td><td>90.1299 %</td><td>7.38 seconds</td></tr>
<tr><td>Multilayer Perceptron</td><td>90.5257 %</td><td>322.78 seconds</td></tr>
<tr><td>Kstar</td><td>94.7747 %</td><td>less than 1 seconds</td></tr>
</table>

We used select attributes in WEKA that chose the best feature vectors for identification information. These feature vectors are min1, min2, min3, max1, max2, max3, sum1, sum2, mean1, mean2, var1, var2, var3, skewness1, skewness2, skewness3. Then we did the tests again with those 5 classifiers. As it shown in table 5.2. Naive Bayes success rate increased around 2% and Kstar success rate increased about 1%. Details for other classifiers also shown in the same table.

TABLE 5.2: Identification Success Rate from Walking Data (2)

Classifier	**Success**	**Time Taken to Build Model**	**Feature**
Naive Bayes	86.3146 %	less than 1 seconds	min1, min2, min3 max1, max2, max3 sum1, sum2 mean1, mean2 var1, var2, var3 skewness1 skewness2,skewness3
Decision Tree (J48)	85.3193 %	0.73 seconds	
Bagging	90.1576 %	4.01 seconds	
Multilayer Perceptron	89.8535 %	287.42 seconds	
Kstar	95.9082 %	less than 1 seconds	

For Kstar and multilayer perceptron, we did tests again using test option 'supplied test set' with feature vectors that were gathered from the walking data (collected from 110 subjects). 60% of total set was chosen for training and 40% was chosen for testing in this method. As it can be seen in table 5.1, test was done by using feature vectors. While Kstar classifier success rate was 94.7747% with cross validation method, it was 82.2754% success rate for Kstar with supplied test set methods. When we did the tests again without using kurtosis1, kurtosis2, kurtosis3 success rate increased to 1,5% with 83.5944% success rate. More details are given in table 5.3 and table 5.4

TABLE 5.3: Identification Success Rate from Walking Data (3)

Features	**min1/2/3, max1/2/3, sum1/2/3, rnge1/2/3, mean1/2/3 var1/2/3, std1/2/3, kurtosis1/2/3, skewness1/2/3 rms1/2/3, resultant**		
Classifier	**Success**	**Time Taken to Build Model**	**Time Taken to Test Model on Supplied Test Set**
Kstar	82.2754 %	less than 1 seconds	157.93 seconds
Multilayer Perceptron	76.0099 %	206.69 seconds	0.14 seconds

TABLE 5.4: Identification Success Rate from Walking Data (4)

Features	**min1/2/3, max1/2/3, sum1/2/3, rnge1/2/3, mean1/2/3, var1/2/3, std1/2/3, skewness1/2/3, rms1/2/3, resultant**		
Classifier	**Success**	**Time Taken to Build Model**	**Time Taken to Test Model on Supplied Test Set**
Kstar	83.5944 %	less than 1 seconds	145.77 seconds
Multilayer Perceptron	77.2465 %	213.5 seconds	0.11 seconds

The walking data was collected by use of Nexus 5 smartphone and Samsung Galaxy Note 10.1 tablet from 34 people. Activities were performed with the mobile devices (smartphone and tablet) at the same time. Subjects carried Nexus 5 in their pocket and they hold the tablet by their hands during the activities. Some success rates are shown in the table 5.5 below. For Naive Bayes, while the success rate was 98.4668% by using the smartphone, the success rate was 88.1803% by using the tablet. There was 10% difference between tablet results and smartphone results. Main reason for this difference is movements and the position of smart devices. We got better success rates with the smartphone because it was replaced in subjects' pockets and there were more motions in there rather than holding the device by hand. If we increase the number of subjects, the success rate would decrease probably. We think if the smart device is replaced in a pocket, we can get better success rates.

Table 5.5: Identification Success Rate from Walking Data (5)

Classifier	**Phone Success**	**Tablet Success**	**Features**
Naive Bayes	98.4668 %	88.1803 %	min1,min2,min3 max1,max2, max3 sum1,sum2,sum3, rnge1, rnge2,rnge3 mean1,mean2,mean3 var1,var2,var3, std1,std2,std3 rms1,rms2,rms3, kurtosis1,kurtosis2,kurtosis3 skewness1,skewness2,skewness3 resultant
DecisionTree (J48)	94.0375 %	84.4262 %	
Bagging	96.4225 %	88.1803 %	
Multilayer Perceptron	99.1482 %	89.9836 %	
Kstar	99.1482 %	93.4262 %	

As mentioned in data collection chapter, with IOS based application, the running data was collected from 90 test subjects. In table 5.6, test result based on using 5 different classifiers and 10-fold cross validation were shown with feature vectors which were from 90 people. Test result, time and used features were also shown in this table. Similar to the previous studies, K star algorithms are the algorithm for identification systems that gives the best results with overall success rate 85.449%. Naive Bayes algorithm that is used for most studies and which gives results very quickly, accomplished 78.7981% success rate. More details which were gathered from WEKA also shown in the table below.

TABLE 5.6: Identification Success Rate from Running Data (1)

Classifier	Success	Time Taken to Build Model	Features
Naive Bayes	78.7981 %	0.02 seconds	min1,min2,min3 max1,max2,max3 sum1,sum2,sum3 rnge1,rnge2,rnge3 mean1,mean2,var1 mean3,var2,var3 std1, std2, std3 kurtosis1,kurtosis2 kurtosis3,skewness1 skewness2,skewness3 rms1, rms2, rms3 resultant
Decision Tree (J48)	76.0297 %	1.18 seconds	
Bagging	82.6806 %	5.46 seconds	
Multilayer Perceptron	83.052 %	213.48 seconds	
Kstar	85.449 %	less than 1 seconds	

We used select attributes in WEKA that chose the best feature vectors for identification information. These feature vectors are min1, min2, min3, max1, max2, max3, sum1, sum2, rngel, mean1, mean2, mean3, var2, skewness1, skewness2, skewness3, rms3. Then we did the tests again with those 5 classifiers. As it shown in table 56. Naive Bayes success rate increased around 2% and Kstar success rate increased about 1,5%. Details for other classifiers also shown in the same table.

TABLE 5.7: Identification Success Rate from Running Data (2)

Classifier	Success	Time Taken to Build Model	Feature
Naive Bayes	80.9926 %	0.05 seconds	min1, min2, min3 max1, max2, max3 sum1,sum2,rnge1 mean1, mean2 mean3, var2, skewness1,skewness2 skewness3, rms3
Decision Tree (J48)	77.1776 %	0.49 seconds	
Bagging	82.1404 %	2.65 seconds	
Multilayer Perceptron	82.8157 %	152.98 seconds	
Kstar	87.002 %	less than 1 seconds	

The running data also was collected by use of Nexus 5 smartphone and Samsung Galaxy Note 10.1 tablet from 34 people. Activities were performed with the mobile devices (smartphone and tablet) at the same time. While the smartphone was place to subjects' pocket, they hold the tablet by their hands. Some success rates are shown in the table below. For Naive Bayes, while the success rate was 89.2857% by using the smartphone, the success rate was 88.1967% for Naive Bayes by using the tablet. The difference was 1% between tablet results and smartphone results this time. Even though this difference was 10% for the walking data, but it was 1% for the running data. Although results for walking data have more success rates than results for running data, success rate results for running data using by tablet are very close to using by smartphone. While the mobile device was held by hand when the running activities were performed, data could be separated from the noise.

TABLE 5.8: Identification Success Rate from Running Data (3)

Classifier	Phone Success	Tablet Success	Features
Naive Bayes	89.2857 %	88.1967 %	min1,min2,min3 max1,max2, max3 sum1,sum2,sum3, rnge1,rnge2,rnge3 mean1,mean2, mean3, var1,var2,var3 std1,std2,std3, rms1,rms2, rms3, kurtosis1,kurtosis2,kurtosis3 skewness1,skewness2,skewness3, resultant
DecisionTree (J48)	85 %	80.6557 %	
Bagging	88.5714 %	87.0164 %	
Multilayer Perceptron	95.3571 %	91 %	
Kstar	94.6429 %	92.4754 %	

We compared walking data which was used for identification success rate and running data which was also used for identification success rate. To get more accurate and realistic results, a number of subjects and feature vectors were fixed. Identification success rate for running and walking that was collected from 90 people was shown in the table below. As it can be seen in this table, results for walking increased approximately 10% more. For example, while results of running activities for naive bayes were 78.7981%, results of walking activities were 86.4307%. It is understandable from this study that walking data gives more successful results rather than running data for identification. Because subjects were performing the running activity while the smartphone was replaced in their pockets, more motions produced more noise. As a result, we can say walking activity while smartphone is replaced in a pocket gives the most accurate data and results.

Table 5.9: Identification Success Rate from Walking and Running Data

Classifier	Walking Data's Success	Running Data's Success	Time Walking	Time Running	Features
Naive Bayes	86.4307 %	78.7981 %	0.04 seconds	0.02 seconds	min1,min2,min3 max1,max2,max3 sum1,sum2,sum3 rnge1,rnge2,rnge3 mean1,mean2,var1 mean3,var2,var3 std1, std2, std3 kurtosis1,kurtosis2 kurtosis3,skewness1 skewness2,skewness3, rms1, rms2, rms3 resultant
Decision Tree(J48)	86.7994 %	76.0297 %	1.05 seconds	1.18 seconds	
Bagging	89.7124 %	82.6806 %	4.72 seconds	5.46 seconds	
MultilayerPerceptron	91.5929%	83.052 %	262.4 seconds	213.48 seconds	
Kstar	95.6121 %	85.449 %	less than 1 seconds	less than 1 seconds	

5.3 Experimental Analysis Results of Age Recognition

In this study, we present age recognition by use of behavioral biometrics in smartphones. We used inbuilt accelerometer sensors of a smartphone to acquired gait data from it. The walking and running data were collected from two different smartphones, Iphone 6 (IOS Platform) and Nexus 5 (Android Platform) for age recognition. As mentioned, extended and basic feature vectors were gathered from this data. We obtained results in five different classes which are Naive Bayes, Decision Tree (J48), Bagging, Multilayer Perceptron, Kstar for classification.

We categorized the subjects into two groups for age recognition; young and adult. In the beginning, we wanted to have more groups rather than two including teenager, young adult etc. however, since most of the subjects are more less in same ages, it was decided to have only two groups for more realistic results. The age range for this experiment was between 15 and 57. Even though, we collected data from kids who is younger than 15 years old, there was too much noise with their data because they move around too much instead of following the instructions. People who are older than 57 had a hard time to perform the activities. So we had to eliminate this data to reach the more accurate conclusion.

We used IOS app to collect walking data for age recognition from 90 subjects. 54 of 90 people were young people and 36 of 90 people were adult. 2341 gait data were gathered from the subjects. 1416 gait data was obtained from young subjects and 925 gait data was from adult subjects. The table shown below was created by using the results based on 10 fold cross-valid.

Table 5.10: Age Recognition Success Rate from Walking Data (1)

Classifier	**Success**	**Time Taken to Build Model**	**Feature**
Naive Bayes	64.1704 %	0.01 seconds	min1, min2, min3 max1, max2, max3 sum1,rnge1,rnge2 rnge3 mean1 mean2,rms1
Decision Tree (J48)	95.0021 %	0.09 seconds	
Bagging	95.6429 %	0.22 seconds	
Multilayer Perceptron	91.0295 %	3.14 seconds	
Kstar	98.6758 %	less than 1 seconds	

Subjects yielded promising results for age recognition. While the success rate was approximately 64% for Naive Bayes, for other 4 classes (Decision Tree (J48), Bagging, Multilayer Perceptron, Kstar) success rate was more than 98%.

TABLE 5.11: Summary for Age Recognition Success Rate from Walking Data (1)

Correctly Classified Instances	2310
Incorrectly Classified Instances	31

TABLE 5.12: Confusion Matrix for Age Recognition Success Rate from Walking Data(1)

	Classified as	
	Young=a	**b=Adult**
a=Young	1403	13
b=Adult	18	907

In the table 5.11 shown below, it can be seen that Correctly Classified Instances and Incorrectly Classified Instances. Table 5.12 has details about female and male sample rates with confusion matrix for Kstar Classification Method

We used Android app to collect walking data for age recognition from 34 subjects. 18 of 90 people were young people and 16 of 34 people were male. 603 gait data were gathered from the subjects. 311 gait data was obtained from young subjects and 292 gait data was from adult subjects. The table shown below was created by using the results based on 10 fold cross-valid.

TABLE 5.13: Age Recognition Success Rate from Walking Data (2)

Classifier	**Success**	**Time Taken to Build Model**	**Feature**
Naive Bayes	76.1194 %	0.02 seconds	min1, min2, min3 max1, max2, max3 sum1,rnge1,rnge2 rnge3 mean1 mean2,rms1
Decision Tree (J48)	95.0249 %	0.06 seconds	
Bagging	94.0299%	0.22 seconds	
Multilayer Perceptron	96.1857 %	0.85 seconds	
Kstar	98.6733 %	less than 1 seconds	

Subjects yielded promising results for age recognition. While the success rate was approximately 76% for Naive Bayes, for other 4 classes (Decision Tree (J48), Bagging, Multilayer Perceptron, Kstar) success rate was more than 98%.

Table 5.14: Summary for Age Recognition Success Rate from Walking Data (2)

Correctly Classified Instances	595
Incorrectly Classified Instances	8

Table 5.15: Confusion Matrix Age Recognition Success Rate from Walking Data (2)

	Classified as	
	Young=a	**b=Adult**
a=Young	307	4
b=Adult	4	307

In the table 5.14 shown below, it can be seen that Correctly Classified Instances and Incorrectly Classified Instances. Table 5.15 has details about female and male sample rates with confusion matrix for Kstar Classification Method

We used IOS app to collect running data for age recognition from 60 subjects. 39 of 60 people were young people and 21 of 60 people were adult. 1670 gait data were gathered from the subjects. 919 gait data was obtained from young subjects and 751 gait data was from adult subjects. The table shown below was created by using the results based on 10 fold cross-valid.

Table 5.16: Age Recognition Success Rate from Running Data (1)

Classifier	**Success**	**Time Taken to Build Model**	**Feature**
Naive Bayes	61.6766 %	0.03 seconds	min1, min2, min3 max1, max2, max3 sum1,rnge1,rnge2 mean1 mean2,rms1
Decision Tree (J48)	87.1856 %	0.2 seconds	
Bagging	90.6587 %	0.32 seconds	
Multilayer Perceptron	85.2096 %	0.81 seconds	
Kstar	95.9281 %	less than 1 seconds	

Table 5.17: Summary for Age Recognition Success Rate from Running Data (1)

Correctly Classified Instances	1602
Incorrectly Classified Instances	68

Table 5.18: Confusion Matrix Age Recognition Success Rate from Running Data (1)

	Classified as	
	Young=a	**b=Adult**
a=Young	886	33
b=Adult	35	716

In the table 5.17 shown below, it can be seen that Correctly Classified Instances and Incorrectly Classified Instances. Table 5.18 has details about female and male sample rates with confusion matrix for Kstar Classification Method.

We used Android app to collect running data for age recognition from 20 subjects. 12 of 20 people were young people and 8 of 20 people were adult. 246 gait data were gathered from the subjects. 144 gait data was obtained from young subjects and 102 gait data was from adult subjects. The table shown below was created by using the results based on 10 fold cross-valid.

TABLE 5.19: Age Recognition Success Rate from Running Data (2)

Classifier	**Success**	**Time Taken to Build Model**	**Feature**
Naive Bayes	78.0488 %	less than 1 seconds	min1, min2, min3 max1, max2, max3 sum1,rnge1,rnge2 rnge3 mean1 mean2,rms1
Decision Tree (J48)	89.4309 %	0.03 seconds	
Bagging	90.2439 %	0.07 seconds	
Multilayer Perceptron	93.0894 %	0.4 seconds	
Kstar	96.3415 %	less than 1 seconds	

TABLE 5.20: Summary for Age Recognition Success Rate from Running Data (2)

Correctly Classified Instances	237
Incorrectly Classified Instances	9

TABLE 5.21: Confusion Matrix Age Recognition Success Rate from Running Data (2)

	Classified as	
	Young=a	**b=Adult**
a=Young	142	2
b=Adult	7	95

In the table 5.20 shown below, it can be seen that Correctly Classified Instances and Incorrectly Classified Instances. Table 5.21 has details about female and male sample rates with confusion matrix for Kstar Classification Method.

5.4 Experimental Analysis Results of Gender Recognition

In this study, we present gender recognition by use of behavioral biometrics in smartphones. We used inbuilt accelerometer sensors of a smartphone to acquired gait data from it. For gender recognition, we investigated Naive Bayes, Decision Tree (J48), Bagging, Multilayer Perceptron, Kstar for classification [49].

Preprocess and segmentation steps for gender recognition were done as it was expressed in chapter 4. For feature extraction, on the other hand, using 7 feature bases we acquired 15 feature vectors. These vectors are; max1, max3, var3, sum1, sum2, sum3, rnge1, rnge2, mean2, mean3, std1, std2, std3, rms2, and rms3. The reason not using some base vectors, they weren't able to recognize genders and not distributing samples to get more accurate results.

We used IOS app to collect walking data for gender recognition from 90 subjects. 35 of 90 people were female and 55 of 90 people were male. 3043 gait data were gathered from the subjects. 1966 gait data was obtained from male subjects and 1077 gait data was from female subjects. The table shown below was created by using the results based on 10 fold cross-valid.

Table 5.22: Gender Recognition Success Rate from Walking Data (1)

Classifier	Success	Time Taken to Build Model	Features
Naive Bayes	74.4003 %	0.01 seconds	max1, max3,var3 sum1,sum2,sum3 rnge1, rnge2 mean2, mean3, std1,std2,std3 rms2, rms3
DecisionTree (J48)	93.6904 %	0.14 seconds	
Bagging	96.2537 %	0.4 seconds	
Multilayer Perceptron	94.0519 %	4.74 seconds	
Kstar	98.5212 %	less than 1 seconds	

Subjects yielded promising results for gender recognition. While the success rate was approximately 74% for Naive Bayes, for other 4 classes (Decision Tree (J48), Bagging, Multilayer Perceptron, Kstar) success rate was more than 90%. Comparing only two different sets (female and male) could be the reason for high success rate. Unfortunately, Naive Bayes didn't give the same high success rate.

Table 5.23: Summary for Gender Recognition Success Rate from Walking Data (1)

Correctly Classified Instances	2264
Incorrectly Classified Instances	779

Table 5.24: Confusion Matrix for Gender Recognition Success Rate from Walking Data (1)

	Classified as	
	Female=a	**b=Male**
a=Female	1512	454
b=Male	325	752

In the table 5.23 shown below, it can be seen that Correctly Classified Instances and Incorrectly Classified Instances. Table 5.24 has details about female and male sample rates with confusion matrix for Naive Bayes Classification Method.

We used Android app to collect walking data for gender recognition from 34 subjects. 15 of 34 people were female and 19 of 34 people were male. 552 gait data were gathered from the subjects. 310 gait data was obtained from male subjects and 242 gait data was from female subjects. The table shown below was created by using the results based on 10 fold cross-valid.

Table 5.25: Gender Recognition Success Rate from Walking Data(2)

Classifier	**Success**	**Time Taken to Build Model**	**Features**
Naive Bayes	88.587 %	less than 1 seconds	max1, max3,var3
DecisionTree (J48)	96.1957 %	0.03 seconds	sum1,sum2,sum3 rnge1, rnge2
Bagging	96.7391 %	0.03 seconds	mean2, mean3,
Multilayer Perceptron	99.0942 %	0.71 seconds	std1,std2,std3
Kstar	99.6377%	less than 1 seconds	rms2, rms3

Subjects yielded promising results for gender recognition. While the success rate was approximately 89% for Naive Bayes, for other 4 classes (Decision Tree (J48), Bagging, Multilayer Perceptron, Kstar) success rate was more than 99%. Comparing only two different sets (female and male) could be the reason for high success rate. Unfortunately, Naive Bayes didn't give the same high success rate.

Table 5.26: Summary for Gender Recognition Success Rate from Walking Data (2)

Correctly Classified Instances	489
Incorrectly Classified Instances	63

TABLE 5.27: Confusion Matrix for Gender Recognition Success Rate from Walking Data (2)

	Classified as	
	Female=a	**b=Male**
a=Female	220	22
b=Male	41	269

In the table 5.26 shown below, it can be seen that Correctly Classified Instances and Incorrectly Classified Instances. Table 5.27 has details about female and male sample rates with confusion matrix for Naive Bayes Classification Method.

We used IOS app to collect running data for gender recognition from 60 subjects. 25 of 60 people were female and 35 of 60 people were male. 2058 gait data were gathered from the subjects. 1252 gait data was obtained from male subjects and 806 gait data was from female subjects. The table shown below was created by using the results based on 10 fold cross-valid.

TABLE 5.28: Gender Recognition Success Rate from Running Data (1)

Classifier	Success	Time Taken to Build Model	Features
Naive Bayes	71.8659 %	0.05 seconds	max1, max3,var3 sum1,sum2,sum3 rnge1, rnge2 mean2, mean3, std1,std2,std3 rms2, rms3
DecisionTree (J48)	93.6904 %	0.18 seconds	
Bagging	94.4121 %	0.39 seconds	
Multilayer Perceptron	93.7804 %	3.23 seconds	
Kstar	98.2993%	less than 1 seconds	

Subjects yielded promising results for gender recognition. While the success rate was approximately 72% for Naive Bayes, for other 4 classes (Decision Tree (J48), Bagging, Multilayer Perceptron, Kstar) success rate was more than 90%. Comparing only two different sets (female and male) could be the reason for high success rate. Unfortunately, Naive Bayes didn't give the same high success rate.

TABLE 5.29: Summary for Gender Recognition Success Rate from Running Data (1)

Correctly Classified Instances	1479
Incorrectly Classified Instances	579

In the table 5.29 shown below, it can be seen that Correctly Classified Instances and Incorrectly Classified Instances. Table 5.30 has details about female and male sample rates with confusion matrix for Naive Bayes Classification Method.

Table 5.30: Confusion Matrix for Gender Recognition Success Rate from Running Data (1)

	Classified as	
	Female=a	**b=Male**
a=Female	903	349
b=Male	230	576

We used Android app to collect running data for gender recognition from 20 subjects. 7 of 20 people were female and 13 of 20 people were male. 280 gait data were gathered from the subjects. 166 gait data was obtained from male subjects and 114 gait data was from female subjects. The table shown below was created by using the results based on 10 fold cross-valid.

Table 5.31: Gender Recognition Success Rate from Running Data (2)

Classifier	**Success**	**Time Taken to Build Model**	**Features**
Naive Bayes	82.8571 %	less than 1 seconds	max1, max3,var3 sum1,sum2,sum3 rnge1, rnge2 mean2, mean3, std1,std2,std3 rms2, rms3
DecisionTree (J48)	93.2143 %	less than 1 seconds	
Bagging	94.2857 %	0.02 seconds	
Multilayer Perceptron	97.1429 %	0.4 seconds	
Kstar	97.5%	less than 1 seconds	

Subjects yielded promising results for gender recognition. While the success rate was approximately 83% for Naive Bayes, for other 4 classes (Decision Tree (J48), Bagging, Multilayer Perceptron, Kstar) success rate was more than 97%. Comparing only two different sets (female and male) could be the reason for high success rate. Unfortunately, Naive Bayes didn't give the same high success rate.

Table 5.32: Summary for Gender Recognition Success Rate from Running Data (2)

Correctly Classified Instances	232
Incorrectly Classified Instances	48

Table 5.33: Confusion Matrix for Gender Recognition Success Rate from Running Data (2)

	Classified as	
	Female=a	**b=Male**
a=Female	91	23
b=Male	25	141

In the table 5.33 shown below, it can be seen that Correctly Classified Instances and Incorrectly Classified Instances. Table 5.32 has details about female and male sample rates with confusion matrix for Naive Bayes Classification Method.

Even though success rate for running was expected to be lower when it was compared with success rate for walking just because running data had more noise, success rates were about the same. The reason for this is distributing male and female rates almost equally.

Chapter 6

Conclusion and Future Work

In this thesis, we proposed the evaluation of performance for identification, gender, and age recognition using behavioral biometrics via the most popular smartphones' accelerometer sensors. For this purpose, we selected the most appropriate methods for every recognition process. Accordingly, we investigated the performance of five classifiers which are Naive Bayes, Bagging, Decision Tree, Multilayer Perceptron, and Kstar. We compared classifiers' results for two activities that are walking and running. Additionally, we also evaluated changes in results on same classifiers by using different feature vectors. Data sets which were used for these performance evaluations have differences for every recognition process and activity (from 20 subjects to 110). As a result, we achieved better results using the walking activity data for all recognition processes. As the classification method, Kstar gave us better results other than the classifiers for every activity and recognition process.

For the future work, we plan to try to provide online personal identification, gender and age recognition at the same time. We want to develop a mobile application tool that will run these three processes through the same platform. To achieve this purpose, with the help of the application we will able to collect data. This application will analyze the data and can give us the accurate recognition result. In addition, we'd like to decrease the number of feature vectors. Our other aim is to find a common set for feature vectors which are less than what we used for every 3 recognition process.

Appendix A

Visual Distributions of Features

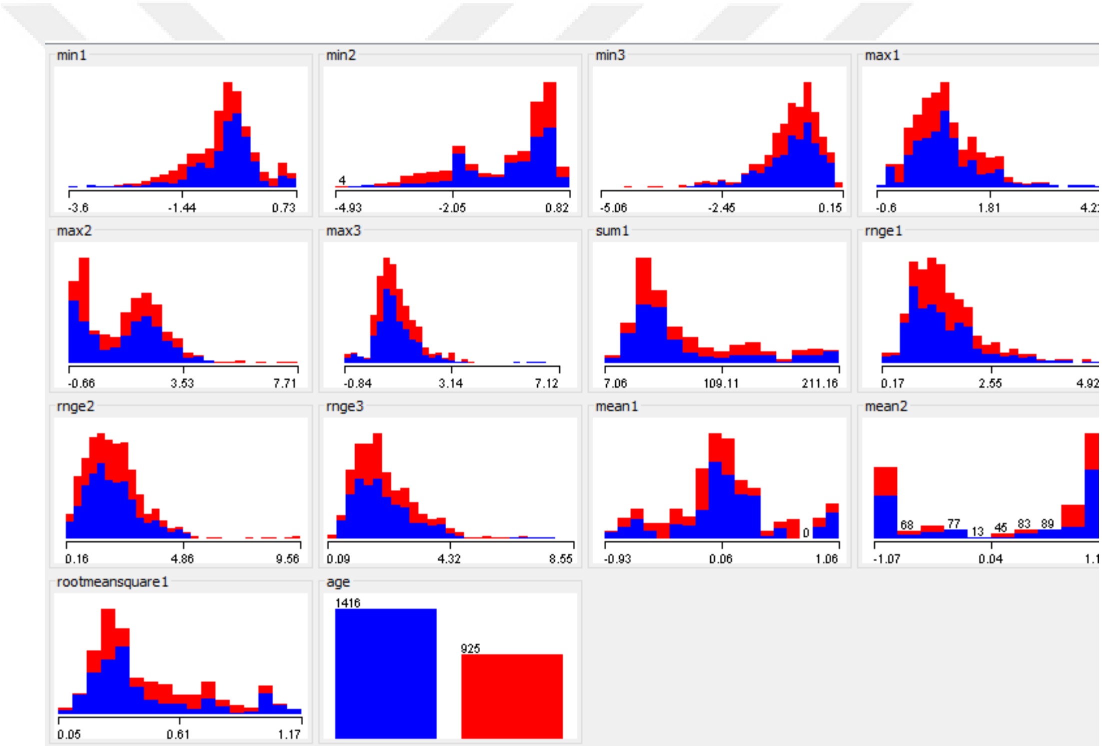

FIGURE A.1: Age Recognition from Walking Data's Features(1)

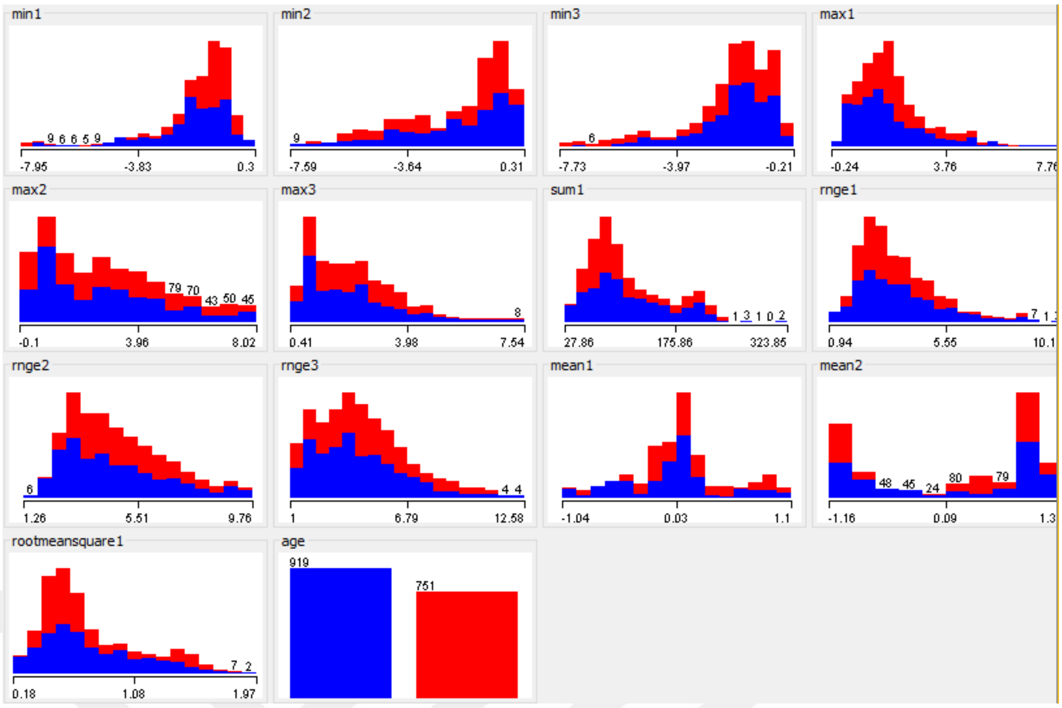

FIGURE A.2: Age Recognition from Running Data's Features(1)

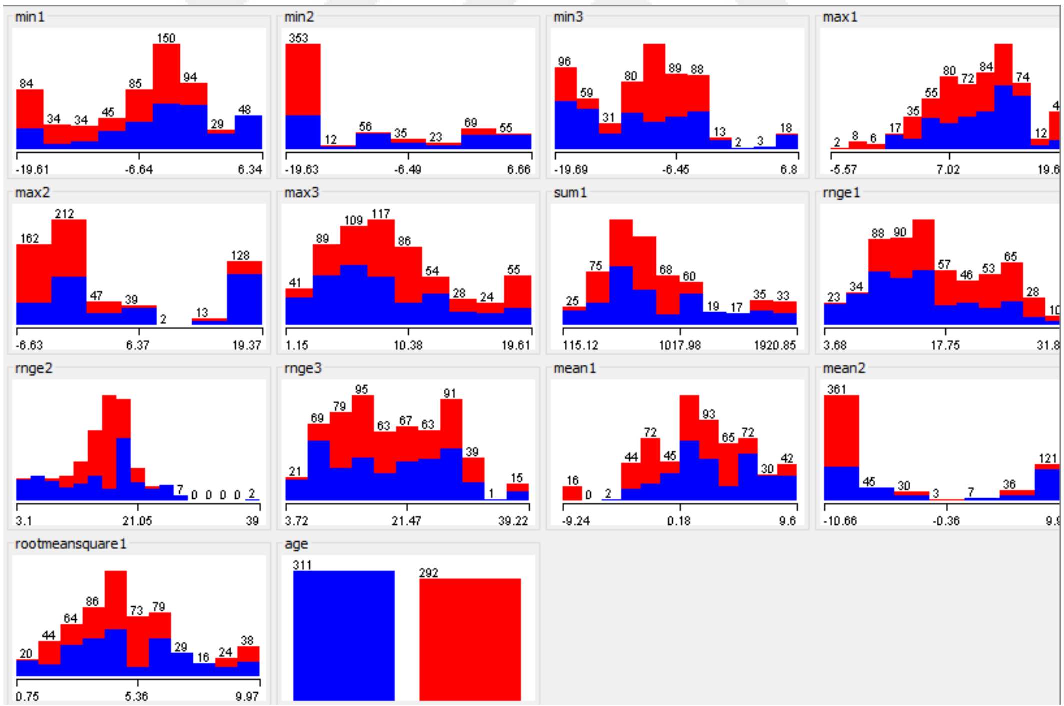

FIGURE A.3: Age Recognition from Walking Data's Features(2)

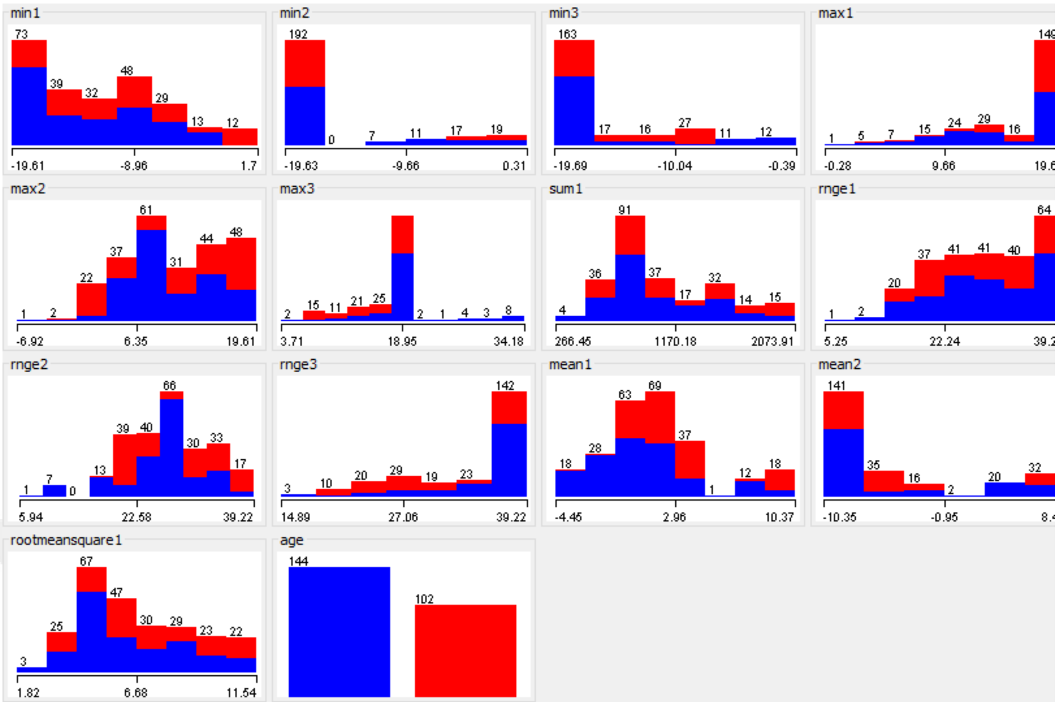

FIGURE A.4: Age Recognition from Running Data's Features(2)

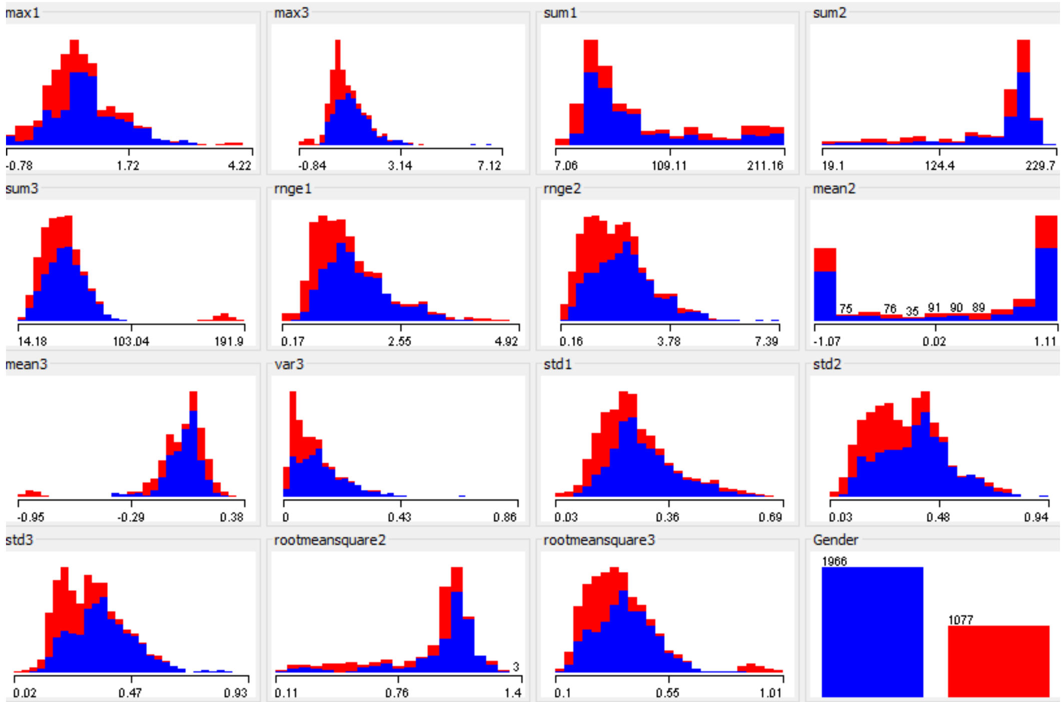

FIGURE A.5: Gender Recognition from Walking Data's Features(1)

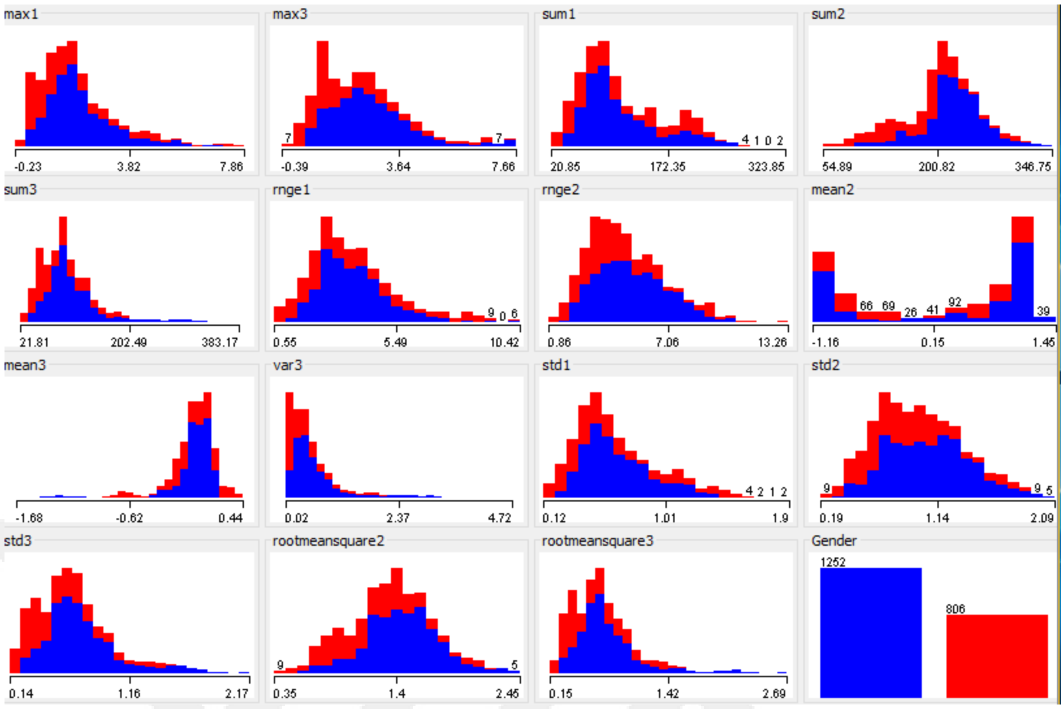

FIGURE A.6: Gender Recognition from Running Data's Features(1)

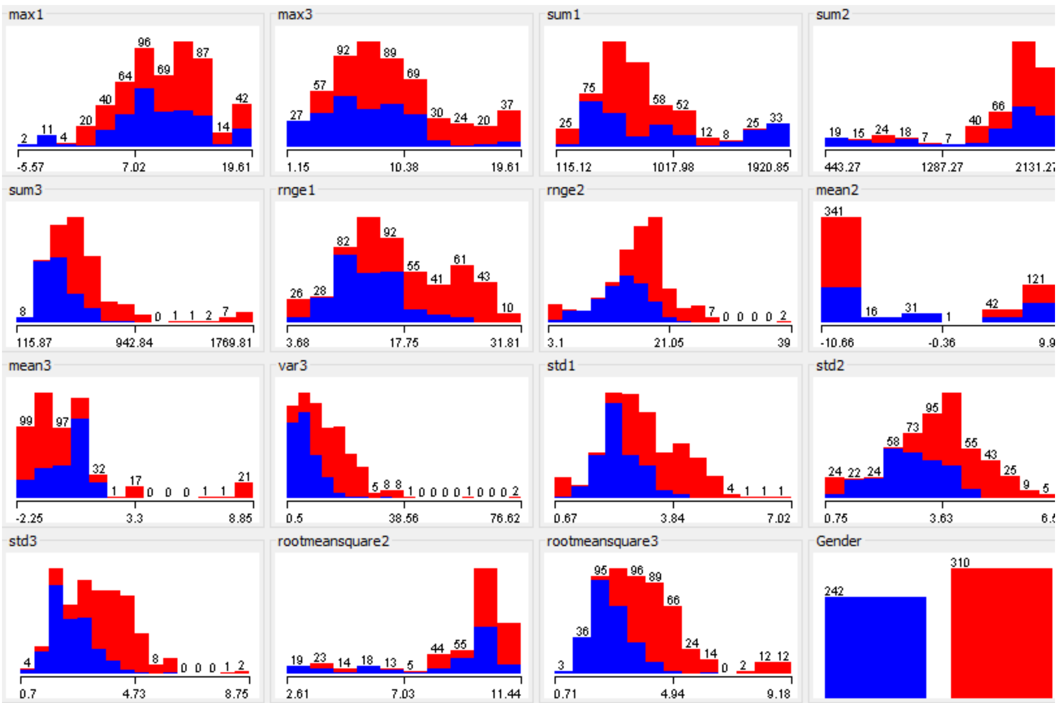

FIGURE A.7: Gender Recognition from Walking Data's Features(2)

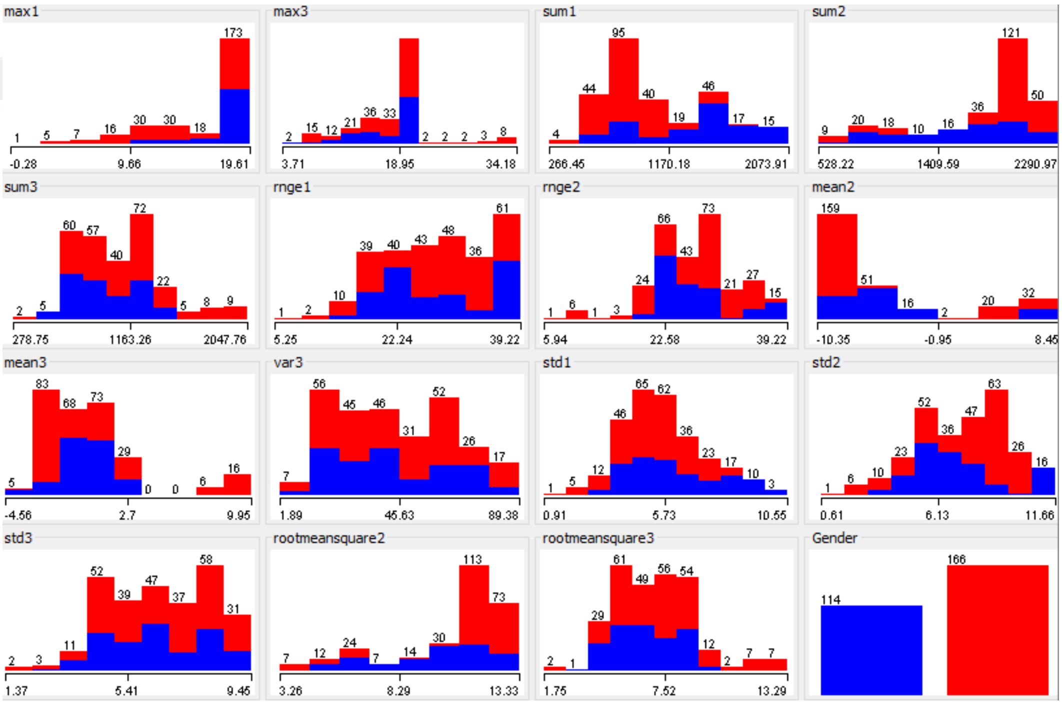

FIGURE A.8: Gender Recognition from Running Data's Features(2)

Bibliography

[1] K. Curran. Biometric authentication: Making mobile devices and apps safer, 2016. URL http://betanews.com/2016/04/04/biometric-authentication-making-mobile-devices-and-apps-safer/.

[2] USTelecom Media. Security developments sentiment on biometric technologies, 2016. URL https://www.ustelecom.org/blog/security-developments-sentiment-biometric-technologies.

[3] The different types of biometrics information technology essay, 2015. URL https://www.ukessays.com/essays/information-technology/the-different-types-of-biometrics-information-technology-essay.php.

[4] The stronger the gmail email password the better is the account security? URL http://www.webdevelopersnotes.com/change-gmail-password.

[5] L. Lorenzetti and Fortune. These were the top 10 selling smartphones in june, 2015. URL http://time.com/3967880/smartphone-sales-june/.

[6] C. Hall. Best smartphones 2016: The best phones available to buy today, 2016. URL http://www.pocket-lint.com/news/120309-best-smartphones-2016-the-best-phones-available-to-buy-today.

[7] A. Mahajan. Worldwide active smartphone users forecast 2014 - 2018: More than 2 billion by 2016, 2014. URL http://dazeinfo.com/2014/12/18/worldwide-smartphone-users-2014-2018-forecast-india-china-usa-report/.

[8] 2 billion consumers worldwide to get smart(phones) by 2016, 2014. URL https://www.emarketer.com/Article/2-Billion-Consumers-Worldwide-Smartphones-by-2016/1011694.

[9] R.Goodrich. Accelerometer vs. gyroscope: What's the difference?, 2013. URL `http://www.livescience.com/40103-accelerometer-vs-gyroscope.html`.

[10] E.Dogan. Efficient feature selection for online activity recognition on smart phones, 2008.

[11] M. Kose. Performance evaluation of classification method for online activity recognition on smart phones, 2009.

[12] C. Randell and H. Muller. Context awareness by analysing accelerometer data. In *The fourth international symposium on Wearable computers: Digest of papers, October 16-17, 2000, Atlanta, Georgia, United States: IEEE Computer Society*. IEEE, 2000.

[13] J. Mantyjarvi, J. Himberg, and T. Seppanen. Recognizing human motion with multiple acceleration sensors. In *IEEE international conference on systems, man and Cybernetics: E-systems and e-man for cybernetics in cyberspace: October 7-10, 2001, Tucson Conventiont center, Tucson, Arizona, USA*. IEEE, 2001.

[14] L.Bao and S.S.Intille. Activity recognition from user-annotated acceleration data. In *Pervasive computing: Second international conference, Linz/Vienna, Austria*. Springer, 2004.

[15] N. Ravi, N. Dandekar, P. Mysore, and M. L. Littman. Activity recognition from accelerometer data. In *17th conference on Innovative applications of artificial intelligence*, 2005.

[16] J. Mantyjarvi, M. Lindholm, E. Vildjiounaite, S.-M. Makela, and H. Ailisto. Identifying people from gait pattern with accelerometers. In *Pervasive computing: Second international conference, Linz/Vienna, Austria*, 2005. URL `http://www.vtt.fi/inf/julkaisut/muut/2005/ICASSP05.pdf`.

[17] M. F. Nowlan. Human identification via gait recognition using accelerometer gyro forces. 2009. URL `http://cswww.cs.yale.edu/homes/mfn3/pub/mfn_gait_id.pdf`.

[18] T. Brezmes, J.-L. Gorricho, and J. Cotrina. Activity recognition from accelerometer data on a mobile phone. In *Distributed computing, artificial intelligence,*

bioinformatics, soft computing, and ambient assisted living 10th international work-conference on artificial neural networks, IWANN 2009 workshops, Salamanca, Spain. Springer, 2009.

[19] S. Sprager and D. Zazula. Gait identification using cumulants of accelerometer data. In *Distributed computing, artificial intelligence, bioinformatics, soft computing, and ambient assisted living 10th international work-conference on artificial neural networks, IWANN 2009 workshops, Salamanca, Spain.* WSEAS, 2009.

[20] J. Kwapisz, G. Weiss, and S. Moore. Cell phone-based biometric identification. In *Fourth IEEE International Conference on Biometrics Compendium.* IEEE, 2010.

[21] A. Rasekh, C.-A. Chen, and Y. Lu. Human activity recognition using smartphone,, 2011. URL `http://arxiv.org/ftp/arxiv/papers/1401/1401.8212.pdf`.

[22] G. M. Weiss and J. W. Lockhart. Identifying user traits by mining smart phone accelerometer data. In *Fifth International Workshop on Knowledge Discovery from Sensor Data*, 2011.

[23] S. Ferrer and T. Ruiz. Science direct travel behavior characterization using raw accelerometer data collected from smartphones. Procedia - Social and Behavioral Sciences, pages 140–149, 2014.

[24] N.Celenli, K. N. Sevis, M. Esgin, K. Altundag, and U. Uludag. An unconstrained activity recognition method using smart phones. In *Biometrics Special Interest Group (BIOSIG)*, pages 1–7. IEEE, 2014.

[25] B. Aguiar, T. Rocha, J. Silva, and I. Sousa. Accelerometer-based fall detection for smartphone. In *Medical Measurements and Applications (MeMeA).* IEEE, 2014.

[26] A. Jain and V. Kanhangad. Exploring orientation and accelerometer sensor data for personal authentication in smartphones using touchscreen gestures. In *Pattern Recognition Letters, Elsevier*, pages 351–360, 2015. URL `http://www.sciencedirect.com/science/article/pii/S0167865515002056`.

[27] V. Osmani, A. Gruenerbl, G. Bahle, C. Haring, P. Lukowicz, and O. Mayora. Detecting depressive and manic episodes. In *IEEE Pervasive Computing Smartphones in Mental Health*, 2015. URL `http://arxiv.org/ftp/arxiv/papers/1510/1510.01665.pdf`.

[28] E. Garcia-Ceja, V. Osmani, and O. Mayora. Automatic stress detection in working environments from smartphones' accelerometer data: A first step. In *Journal of Biomedical and Health Informatics*. 2015. URL http://arxiv.org/pdf/1510.04221.pdf.

[29] Ruben, Juan, Roberto, Fernando, and Jose. Feature extraction from smartphone inertial signals for human activity segmentation. In *Journal of Biomedical and Health Informatics*. ACM, 2016.

[30] A.Laghari, W.Rehman, and Z.A.Memon. Biometric authentication technique using smartphone sensor. In *Applied Sciences and Technology (IBCAST), 2016 13th International Bhurban Conference on*. IEEE, 2016.

[31] J. Yang. Toward physical activity diary: Motion recognition using simple acceleration features with mobile phones. In *In Proceedings of the 1st International Workshop on Interactive Multimedia for Consumer Electronics*, 2009.

[32] L. Sun, D. Zhang, B. Li, B. Guo, and S. Li. Activity recognition on an accelerometer embedded mobile phone with varying positions and orientations. In *In Ubiquitous Intelligence and Computing*, 2010.

[33] J. R. Kwapisz, G. M. Weiss, and S. A. Moore. Activity recognition using cell phone accelerometers. 2010.

[34] A. Henpraserttae, S. Thiemjarus, and S. Marukatat. Accurate activity recognition using a mobile phone regardless of device orientation and location. In *Body Sensor Networks (BSN), 2011 International Conference on*. IEEE, 2011.

[35] S. Dernbach, B. Das, N. C. Krishnan, B. L. Thomas, and D. J. Cook. Simple and complex activity recognition through smart phones. In *In Intelligent Environments (IE), 2012 8th International Conference*, 2012.

[36] Y. E. Ustev, O. Durmaz Incel, and C. Ersoy. User, device and orientation independent human activity recognition on mobile phones: Challenges and a proposal. 2013.

[37] M. T. Shrove. Sensors smartphone developments. URL http://slideplayer.com/slide/10707486/.

[38] Motion sensors, . URL https://developer.android.com/guide/topics/sensors/sensors_motion.html.

[39] Sensors Media Management. Android development. URL http://www.ce.unipr.it/~picone/mobdev/Unipr_AD_10.pdf.

[40] Sensors overview, . URL https://stuff.mit.edu/afs/sipb/project/android/docs/guide/topics/sensors/sensors_overview.html.

[41] Charmaine Jirgens. How to use the accelerometer on an android mobile device.

[42] Andrdoid sensors overview. URL http://www.th7.cn/Program/Android/201608/938553.shtml.

[43] Apple. core motion framework reference, . URL http://cdn.cocimg.com/cms/uploads/soft/130911/4196-130911100236.pdf.

[44] Cmmotionmanager class reference, . URL http://www.apihome.cn/api/ios/CMMotionManager%20Class%20Reference.html.

[45] Jason Brownlee. How to prepare data for machine learning.

[46] Y.E. Usrev. User, device, orientation and position independent human activity recognition on smart phones, 2008.

[47] Mathworks documentation. URL https://se.mathworks.com/.

[48] Data mining software in java. URL http://www.cs.waikato.ac.nz/ml/weka/.

[49] A. Jain and V. Kanhangad. Investigating gender recognition in smartphones using accelerometer and gyroscope sensor readings. In *International Conference on Computational Techniques in Information and Communication Technologies*, 2016.