Biometric Identification and Authentication Using Time Series Classification for Mouse and Eye Movements

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

by

Fedaa ELDERDESAWE

in partial fulfillment for the
degree of Master of Science

in
Industrial and Systems Engineering
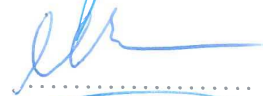
İSTANBUL
ŞEHİR
UNIVERSITY

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 Industrial and Systems Engineering .

APPROVED BY:

Asst. Prof. Dr. Mustafa Gokce Baydogan ....
(Thesis Advisor)

Asst. Prof. Dr. Mehmet Baysan ....

Assoc. Prof. Dr. Ali Fuat Alkaya ....

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: 25 October 2017

SEAL/SIGNATURE:

Declaration of Authorship

I, Fedaa ELDERDESAWE, declare that this thesis titled, 'Biometric Identification and Authentication Using Time Series Classification for Mouse and Eye Movements ' 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:

Date: 25/10/2017

Biometric Identification and Authentication Using Time Series Classification for Mouse and Eye Movements

Fedaa ELDERDESAWE

Abstract

Security plays a very important role in modern world where almost everything is done with the computer. It is agreed that biometric recognition systems require the combined analysis of multiple behavioral traits or physiological characteristics. In addition, those systems are considered to be the most flexible and effective mode of identifying and authenticating individuals as the person does not need to remember any password, or carry smart cards.

The human body can remember the movement of mouse and the gaze if that action is practiced a lot, which mean when the user want to be authenticated in to computer so he will not forget the mouse and eye actions. So, this actions can be utilized in way of password authentication system, in which if the user implement the right movements can be considered as an authenticated user. Otherwise the system will reject the user. So for experimenting the authentication system, Different time series datasets consisting of mouse movements and gaze positions were analyzed and an authentication model was developed. It is shown that the users can be authenticated by proving their claimed identities using the developed model. This thesis investigates mouse and eye coordinates for user recognition scheme that introduce a random forest classification model for Mouse and eye movements to recognize these movements. The focus of this thesis is on the classification methods of time series, including similarity measures and random forest . Features are extracted from the mouse and eye movements raw data and implementing 1.Nearest Neighbor and Random forest to classify users. The accuracy of the identification varies with the variety of features used The experimental results were competing with our proposed biometric authentication model. The accuracy achieved by 1.Nearest neighbor was not sufficient in predicting users identities by mouse and eye tracking . On the other hand the maximum accuracy from implementing random forest model was 60 % which quietly good in terms of biometric but it is still need development to have perfect biometric model with higher accuracy.

Keywords: Biometric, Time series, Classification, Authentication, Gaze Direction, Mouse Dynamics, Identification, K-nearest Neighbor Classification, Random Forest, Feature Extraction

Fare ve Göz Hareketlerinin Zaman Serileri Sınıflandırma Yöntemleri Kullanılarak Biometrik Tanıma ve Kimlik Doğrulamada Kullanılması

Fedaa ELDERDESAWE

Öz

Güvenlik, neredeyse her işlemin bilgisayar vasıtasıyla gerçekleştirildiği modern dünyada çok önemli bir rol oynamaktadır. Güvenlik alanında kullanılan biyometrik uygulamalara ait ihtiyaçların karşılanmasına yönelik olarak, insan-bilgisayar etkileşimi ve ilgili diğer alanlarda biyometrik araştırmalar hızlı biçimde artış göstermektedir. Biyometrik tanıma sistemleri, çeşitli davranışsal eğilimler veya psikolojik özelliklerin birlikte analizini gerektirmektedir. Ayrıca bu sistemler, en esnek ve en etkili birey tanıma ve kimlik doğrulama biçimi olarak kabul edilmektedir. Bu yolla kişinin herhangi bir şifre hatırlaması veya akıllı kart taşıması gerekmemektedir.

İnsan çok fazla tekrarlanması durumunda fare hareketlerini ve bakışı hatırlayabilmektedir; bir kullanıcıdan bilgisayarda kimliğini doğrulatması istendiğinde kullanıcı fare ve göz hareketlerini unutmayacaktır. Bu nedenle bu eylemler, şifre doğrulama sistemi olarak kullanılabilir. Bu sistemde kullanıcı doğru hareketleri gerçekleştirdiğinde kimliği doğrulanmış bir kullanıcı olarak kabul edilebilir. Aksi durumda sistem kullanıcıyı reddedecektir.

Bu nedenle, kimlik doğrulama sistemini denemek amacıyla amacıyla fare hareketlerinden ve bakış pozisyonlarından oluşan farklı zaman serilerini içeren veri setleri analiz edilerek bir kimlik doğrulama modeli geliştirilmiştir. Kullanıcıların geliştirilen modeli kullanarak, talep edilen kimliği kanıtladıkları ve bu yolla kimliklerinin doğrulandığı görülmüştür. Bu tez çalışması, rastgele karar ağacı algoritması kullanılarak oluşturulan kullanıcı tanıma şeması için fare hareketi ve göz hareket koordinatlarını araştırmaktadır.

Bu tezin odaklandığı nokta, zaman serilerinin sınıflandırılması amacıyla benzerlik ölçümleri ve rastgele karar ağacı algoritmalarını kullanan metodlardan oluşmaktadır.

Özellikler; Fare ve göz hareketlerine ait ham veriler ve kullanıcıları sınıflandırmak için en yakın komşu ve rastgele karar ağacı algoritmalarını kullanan birinci uygulamadan elde edilmiştir. Kullanılan çeşitli unsurlara göre kimlik doğrulama çeşitlilik göstermektedir. Deneysel sonuçlar, önerdiğimiz biyometrik kimlik doğrulama modelindekilerle uyuşmaktadır.

DEDICATION

To my mother soul...

To my brother soul..

To all martyrs I want dedicate my work.

Allah bless them all .

Acknowledgments

I would like to express my most valuable appreciation to my advisor Asst. Prof. Mustafa Gokce Baydogan for his support, contribution, patience, and guidance. Without his patience this thesis would not have been possible. He always direct me to the right direction .

I am also grateful to my other family members and friends for their support and encouragement throughout my study.

Contents

List of Figures

List of Tables

Abbreviations

PIN	**P**ersonal **I** dentification **N**umber
DTW	**D**ynamic **T**ime **W**arping
TSDM	**T**ime **S**eries **D**ata **M**ining
KNN	**K** -**N**earest **N** Neighbor
1NN	**1** **N**earest **N** Neighbor
ED	**E**uclidean **D**istance
SVM	**S**upport **V**ector **M**achine
RF	**R**andom **F**orest
KDD	**K**Knowledge **D**iscovery in **D** atabases
SLT	**S**tatistical **L**earning **T**heory

Chapter 1

Introduction

In the modern world, security plays a very important role as almost most of the critical operations are performed with the computer. In that sense, reliable authentication systems are needed for security purposes. User authentication is an important security solution in information systems especially on a web-based or corporate network [1]. Generally, authentication systems achieve their objective through different factors which can be categorized as follows:

- An object the user has, like an identity card or a security token,

- A thing the person memorize, like a password; a personal identification number (PIN) or a pass phrase,

- A thing belongs to the user, like fingerprints or retina [2].

To increase the security of computers , various software models have been created to simplify the first two category. Password or a smart cards can be easily stolen and users cannot remember it. Moreover, passwords have expiration date which require the setting of a new password on a regular basis. This needs more work and time. Also, there is always uncertainty in these software models as the system cannot guarantee that the authenticated user is the real person who has the authority. To resolve the potential problems with the existing authentication strategies, biometrics (i.e. third category) are considered as an alternative solution for the recognition of the individual. The main motivation behind these efforts is the unreliable nature of the simple username-password authentication or identity cards when compared to complex authentication

like fingerprint or retina scan. This is related to understanding of metrics related to human behavior which is shown to be safer. Biometric identification is known as the method of verifying the precise identity by recognizing one or more physical or behavioral features. It is considered to be one of the future vital approaches for user authentication [3]. Biometric features or biometric identifiers are two main parts : Physiological and behavioral features. A physiological biometric authentication system recognize users using their fingerprints, iris scans, retina scans, facial recognition, hand geometry, wrist veins, palm topology, thermal images voice recognition and DNA tests. The behavioral features include voice prints, handwritten, mouse movements signatures and keystroke dynamics [4].

With the technological advances, the biometric authentication systems have become more and more trustworthy and appropriate, but some challenges and disadvantages still should be taken into account when applied for authentication. Most of the biometric authentication systems are not 100% accurate. Another drawback in biometric systems is fraud attacks or artificial imitation of biometrics. In other words, many of bio-features can possibly be stolen or copied by opponents. For example, fingerprint can be copied as people put their fingerprint whenever they touch any tangible thing object and that make a way of cheating, and voice pattern can be imitated using many created devices which can record secretly a persons voice. In addition, some facial recognition systems can be deceived by an appropriate sized photo of an authorized user. Huge data of gestures like images could produce problems in biometric identification and recognition, as a very complicated photo normally consists of a large number of patterns, that produce a huge data set which requires more time for processing and complicated computation for user identity verification.

To avoid from aforementioned potential problems with the physical features, researchers focused on alternative biometrics based on behavioral features for reliable identification of individuals. The ideal biometric must be continuous over time, easy to implement, cheap, unique, global and highly accepted from the user.

Considering this fact, behavioral authentication approaches consider how people perform certain tasks such as use of pointing devices, mice or touch-pads. Those methods can identify users according to their: typing stroke, mouse dynamics, signatures and etc. Most of the recent biometric identification applications deal with the eye and mouse

movements because of the affordability and simplicity of the data collection. Moreover, it has been shown by many studies that the eye-hand coordination, the ability of performing activities using eye and hand at the same time, has significant potential for identification. The temporal dynamics inherent in eye and hand movements are shown to provide important information regarding the characteristics of individuals. In today's world, where many vital tasks are done with few clicks, the need for reliable, cheap, security systems is growing. Hence, utilization of eye and mouse movements as a gate to biometric authentication for humans may offer certain notable advantages in visual tasks such as reading, exploration of digital displays and in online security.

Many daily tasks are performed using the coordination of eyes and hands and researchers study these movements widely in many scientific fields such as: cognitive science, psychology, medicine and etc. Our eyes and hands move in coordination to execute many of our daily tasks and the temporal dynamics of these movements have been widely studied in cognitive science usability testing and psychology. Recently, this information is being used for user authentication purposes. The main motivation behind the use of mouse movements for identification is that it is almost common to all personal computers. There is no need for a special hardware which makes it cheaper compared to some biometrics like fingerprints. Although capturing of eye movements require camera and software, they are relatively cheap compared to devices such as fingerprint scanners. Moreover, the amount of data to be stored is considerably smaller than those of images.

This research focuses on a certain authentication approach, PIN-Pass. It is based on entering four digits PIN sequence, in which each two successive digits must be different. To login to the computer system, the user should enter the four digit number by clicking them in the given order as illustrated in Figure 1.1. After that, the records of eye movements are collected, the next step is how to extract identification features from these movements which seems to be very hard step and need much study and experiments, in which at the end lead to human identification.

Eye movements are tracked by an eye tracking software which collect the data related to the x and y coordinates of the point a user looking at over time (i.e. gaze). There are two aspects of eye movements which could be analyzed: behavioral and physiological. The physiological aspects are easy to track and requires extraction of the physiological attributes of the person (i.e. muscle movements). On the other hand, the behavioral

aspects focus on the brain activities which force the eye to move. Many researchers studied the eye as a biometric tool for people identification and focused on eye movements as as the physiological part because it easier to analyze. Therefore, eye tracking software finds the exact point of user look which can be defined as "The exact point where the user is looking in period of time" [5].
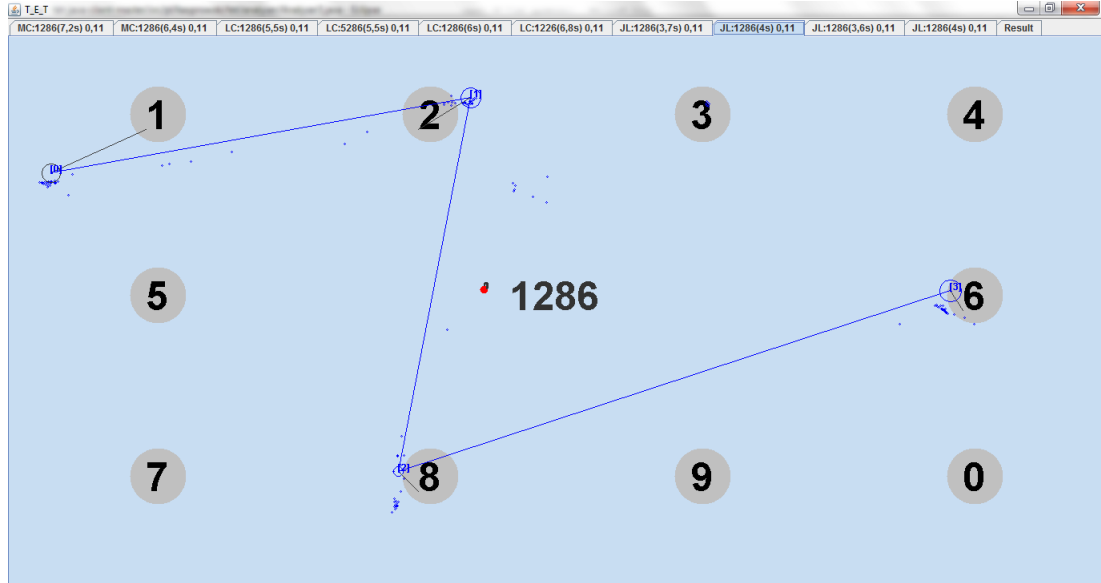
FIGURE 1.1: Example of a screen for eye and mouse tracking experiment

Identification based on only eye movement data has been shown to be inferior and researchers generally combine the eye movements data with additional qualitative data. PIN-Pass also considers mouse movement information to have more accurate authentication model. As an additional information source, mouse movements are considered as: "a behavioral biometric which analyze the behavior data from pointing devices (mouse)" , which shows the possibility of user identification. Compared to eye tracking, capturing mouse dynamics is relatively cheaper as there is no need for a technology to collect the data. Many research studies focus on mouse dynamics as a monitoring tool to follow the user way of moving the mouse to utilize that for authentication. In this study, mouse movements are obtained based on its position in x and y coordinates, mouse click position and time required to finish those actions. The eye gaze and mouse positions are generally recorded as the person should look at the place while clicking the mouse.

The problem of user identification based on the eye and mouse movement data can be formulated as a supervised learning problem. Based on the previously collected data from the individuals, an individual can be identified after performing the required task

provided by PIN-Pass. However working with the type of data from these applications is not straightforward. Firstly, there are two information sources (i.e. mouse and eye movements) and fusion of the information from both sources are required. Moreover, the collected data is simply the gaze and mouse positions over time and this restricts direct application of supervised learning approaches on the collected data. This is an example of *multivariate* time series where variables can be considered as x and y positions over time. The temporal dynamics inherited in the multivariate time series should be taken into consideration to infer information regarding the user. User authentication problem is simply a multivariate times series classification application and there are several studies on the classification of multivariate time series data which mostly focus on sensor data as the time between observations is constant (i.e. regular intervals). The collected data in these experiments is simple irregularly spaced time series observations which challenges the analysis. Moreover, most of the existing studies assume that the time series from each trial have the same length to simplify the analysis.

The algorithms proposed to solve the multivariate classification problem mostly fall into two main categories: distance-based approaches and feature-based approaches. In the distance-based approaches, univariate time-series similarity approaches are modified to make them work for multivariate time series. In the literature, Dynamic Time Warping(DTW)([6]) distance has been commonly used as the benchmark because of its high accuracy in many univariate time series classification problems. This makes researchers focus on extensions of DTW for multivariate setting. Distance-based approaches focus on each variable of multivariate time series individually (i.e. as an independent univariate time series) and require scaling of the variables. Hence, they are prone to loss of important data loss since MTS are not only defined by separate variables but also by relationships between them. Moreover, high-dimensionality introduced by multiple variables and long time series are problematic for distance-based approaches. An efficient authentication system should provide a fast response however classification requires time consuming distance computations for the comparison of the new time series to the existing ones in the database.

To avoid from potential problems of the distance-based approaches, feature-based methods aim at obtaining a rectangular representation of the time series so that they can be provided as an input to any supervised learning algorithm. This way, prediction can be performed with a model which is faster. Simple feature-based approaches extract

information regarding the statistical properties of the variables of multivariate time series such as minimum, maximum, mean, standard deviation and etc. The approaches dealing with the movement data in the biometric identification domain mainly focus on the extraction of the behavioral features such as velocity and acceleration for mouse and eye movements. After feature extraction, classification algorithms are trained on the new representation.

Classification approach is the best approach for analyzing the time series data since many class subjects are available and the need is assigning the subject label for each time series. After that, allocating the important features to identify subjects from others which helps in building good model. There are many classification algorithms that can be used for generating classification models. In this research, the nearest neighbor and random forest algorithms are prefered. Firstly, 1NN approach is used for subject identification which is considered one of the simplest classification methods in data mining and measured by the distance function. The function that is used in this study is Dynamic Time Warping (DTW) and it is explained in details in Chapter 2. Secondly, the biometric identification problem is also approached by extracting the behavioral features related to the mouse and eye movements records of different subjects such as velocity and acceleration of mouse and eye movements. Then, in the verification step, the extracted features are fed to classification random forest model for the identification process. The primary challenge in a biometric recognition system is to find feature representation pattern and similarity measure to minimize the recognition errors of authenticated user to enhance the security of the computer systems. In this study, the primary objective is to evaluate the utility of the KNN method and random forest for identification process using the mouse and eye movements data. However, building distance matrices and constructing features are considered as the most challenging in time consuming process and memory. Moreover, the model is build according to labeled data while the challenge is to use the resulted model and the unlabeled data to identify the unlabeled subjects. This research handles different three data sets, which is limited number for getting high accuracy however the results are still competitive.

The structure of the thesis is as follows:

- Chapter 1 introduces the biometric authentication and the advantages and disadvantages of biometric.

- Chapter 2 talk about biometric authentication applications, time series classification algorithms and time series data mining.

- Chapter 3 introduces the related literature in biometric authentications system

- Chapter 4 explains the data and the steps taken to preprocessing the data

- Chapter 5 summarizes the computational experiments and results

- Chapter 6 concludes by summarizing the results of the research and discusses the potential future work.

Chapter 2

Background

2.1 Biometric Authentication Applications

There is a need for secure and trustful world to feel peaceful. However, the reality is different as there are crimes, computers hackers, and fraud. Biometric authentication comes as a solution for these problems to secure day-life transactions. For a long time, many researchers worked on authentication approaches to prove the identity of people they claim to be. For this purpose, various applications have been invented:

- Fingerprint Authentication: fingerprint authentication has been used for along time in verification the identity of people because fingerprints can not be changed, can not be forgotten, unique to each person, and it eliminates identity mistakes. So, it has been applied in many fields like online services, online banking transactions, customer financial information, health care, mobile banking and secure airlines.

- Voice Authentication: voice biometric authentication is most secure authentication method. Some companies invent vocal-password application to identify users from their voice during their daily transactions using mobile devices or the web.

- E-Commerce Applications: developers study the use of biometric for customer verification and authentication using smart cards and smart phones for Point of Sale (POS) purchases and online shopping.

- Biometrics for Mobile Banking: the traditional way of financial transactions was done face to face. But with biometric technology, it can be done from a distance using computers and the internet such as authentication through ATM machines.

- Multifactor Biometric Authentication: organizations started to benefit from multiple biometric methods such as face, fingerprint, voice, and retina for authentication systems. Wells Fargo, a financial institution, uses the voice with face biometric for customer authentication using mobile devices which makes it impossible for fraud or cheating. Others use photos or video recordings for person authentication.

Properly many research papers are conducted about biometric and security, especially based on mouse and eye movements, and others focused on online banking, email or other user's daily tasks which require reliable, non-disruptive and fast authentication. Some examples include Google applications, financial and health care applications, computer games, office 365 and Drop-box, etc. Apple invented creative applications for eye movements using eye tracking technology by detecting where the user is looking and allow him to cross the device interface [7].

2.2   Eye Movement and Mouse Biometrics

Biometric is known as the method of identifying people by recognizing one or more physicals or behavioral features for those users . Probably, it is one of the future vital approaches for supplying authentication [3]. The great deal and the highest priority of this research is the user authentication which is a method to prove the identity of users by their measurable human characteristics which is a great source of biometric information that can be used to establish or verify a precise identity. User authentication is main security solution in information system especially in being corporate network or a web-based [1]. Generally, authentication and recognition systems achieve their objective through different factors which can be categorized as follows: the user what has like identity card, the user what memorize like a password or a pass phrase and what belongs to user like his fingerprints or retina [2]. The oldest known methods of user authentication are passwords or smart cards authentication. Password or a smart cards can be easily stolen and users cant remember it. On other hand, passwords have expiring date which require reset a new password which need more work and time. In that case there is no

certainty if the authenticated user is the real person who has the authority. That leads to take biometrics in consideration because it is introduced as a key for the previous issues. This research describes further development and evaluation of the authentication system and focuses on biometric authentication which is: "the measurement of physiological or behavioral features that identify and authenticate an individual". Also it can be defined as "it is a process of recognizing persons based on their physical features behavioral and/or learned traits, like face, hand geometry and fingerprint, or behavioral characters, such as signatures and typing stroke".

Eye is one of the most critical parts in the human body. Identification of a person eye is known by following the point on the screen using eye-tracking software. There are two sides of eye movements which could be analyzed; behavioral and physiological aspects, the physiological aspects are easy to track and extract the required attributes of the person because it compiles physiological (muscles), while the behavioral aspects focus on the brain activities which forces the eye to move. Many researches considered the eye as a biometric tool for user identification and most of them focused on the physiological part of the eye movements because it is easier to analyze.

One interesting example for behavioral biometrics is the eye-hand coordination which is the ability of doing activities using eye and hand at the same time like when a user types on the keyboard, visual information send to the brain in order to guide the hand to not make mistakes. Another example for hand-eye coordination is driving as the driver needs the visual information to use his hands for moving the wheel to avoid the accidents. This study concerns with mouse and eye dynamics authentication which counted to behavioral authentication type. The idea is to authenticate people based on the way of how they do things, such as the way they behave using pointing devices (i.e. mice or touch-pads). Those authentication methods identify users according to their typing stroke, mouse dynamics, signatures, gait and voice.

Mouse movements can be defined as the way of user interact with computer through the mouse. Also they are considered as behavioral biometric which analyze the behavior data from pointing devices like mouse, which provide the possibility of user identification based on the user's behavior through using the mouse. Mouse dynamics is a cheap and non-disruptive method as there is no need for a particular hardware for data capturing.

Recently, many researchers have focused on eye and mouse dynamics as a monitoring tool which follow the user way of moving the mouse for authentication. There is a need to capture eye and mouse features through user interacting with computer system in these applications. In this study, eye and mouse movements are obtained based on their position in x and y coordinates, mouse click position, and time required to finish those actions. Unlike other tracking methods, recording mouse actions is very simple and low-cost to implement. The aim is to build accurate user authentication model depending on combining eye and mouse movements so that a supervised learning algorithm can be applied to identify the users from each others. In order to let user access to computer system, mouse movements should be in a certain sequence to select the right numbers on the screen as in Figure 1.1 and match the stored data.

Biometric has several advantages compared with traditional authentication as it is considered the most effective and secure method. The aim of presenting mouse eye authentication system is to capture these advantages. It also improves customer experience and can not be forgotten or lost. Eye movements are an fascinating and likely behavioral biometric for classification. So utilization of eye with mouse movements as a gate to biometric authentication for human show many benefits in visible tasks like reading, online security and discovering a digital displays.

Many researchers succeeded in recognizing and learning the behaviors of the users from their tasks with mouse and eye. Another benefits from using eye movement as a way of authentication is that its data amount of biometric signals is one-dimension which is less than the signal dimensions of images, but still have good data for the subjects recognition . Beside that, eye movements gives an accurate measure of where user visual attention is directed which help in designing websites and ads.

2.3   Time Series Data

A huge data are collected daily in the form of time series in most scientific experiments measurements, which performed over time. These measured values lead to ordered data called time series.

Since our data set is special kind of data which has multiple ordered sequence, that has been studied in this research is a multivariate time series. Those time series consist of

sequential data based on multiple features. It is a general method to build models of all features simultaneously. A multivariate time series X can be clarified by the following formula .

$$x_t = (x_{1t}, x_{2t}, \ldots, x_{nt})^T, \tag{2.1}$$

where (t) is a time index, T sample times and n time series.

Recently research areas are being developed and trended about time series analysis. There are a huge increase of interests in time series methods and algorithms. Companies and analysts are often concerned with discovering patterns in time, for example predicting future patterns. That concerns can be explained by the various applications which produce time data. Practically every single data is changing over time such as data gathered from human, natural and biological processes.

The study of time series give birth to series of data mining challenges in different research areas like pattern recognition, telecommunication data, signal processing, statistics, bank transactions, economics, control engineering, astronomy, meteorology, scientific applications, the volume of product sales, the consumer price index , entertainment and so on.

2.3.1   Time Series Data Mining (TSDM)

The study of time series in statistics has a long history. It is popular that time series analysis is a fundamental to engineering, scientific and business endeavors associated with extracting useful patterns, meaningful statistical information, time series structure principles and predicting the future event. Time series data mining (TSDM) contains five procedures which are collecting, storing, organizing, analyzing, and presenting time series data. It is also theoretical justification in the theory of nonlinear dynamics [8]. The most critical challenges in time series data mining are processing high dimensional time series, being extremely hard and expensive, comparing multiple time series which are time moved or uniformed through amplitude. These make it hard to define similarity measure for huge datasets.

Preprocessing data will participate in handling such problems by normalization and scaling data. Also by data representation and reduction will reduce the huge data into manageable one that could be easily analyzed. Time series segmentation and indexing are considered as preprocessing data mining tasks, which include four major tasks: clustering, indexing, classification and segmentation.

2.3.2   Time series Data Mining Tasks

- **Clustering**

  Clustering for unlabeled data is considered as an essential procedure in the pattern discovery approach. Data clustering is a branch of data mining field which is a process interested in incorporating techniques for finding natural groups, called clusters, in data base. Those groups are needed to be similar or homogeneous groups by maximizing variance between groups while minimizing the variance within the groups [9]. In machine learning field, clustering is considered as unsupervised or semi supervised learning algorithm relying on whether having active parameters or not.

  Clustering of time series is a special solution for data flows. It is applied by using trimmed data representations, Auto-Regressive (AR) models, k means, and efficiency k-center clustering.

- **Motif Discovery**

  Recently, motif mining in time series is the most growing knowledge discovery which referred to detection of recurrently appearing patterns, outliers and novelties or weirdo in a time series data. Novelties are pointed to abnormalities, or weird patterns. This approach was discovered from gene analysis in bioinformatics.

  Many researchers noticed that motif discovery could be a sub routine procedure to discover valuable clusters. Also others proposed random projection algorithm for DNA sequences in order to be more efficient in detecting the anomalies or motifs [10].

- **Segmentation** Segmentation is an important step for time series which partitions time series into discrete classes, and it is performed to many time series fields.

  The idea of segmentation comes from the trying efforts to solve the problem of high dimensional time series by reducing those dimensions through the accurate

approximating of time series while maintaining necessary patterns and characteristics of the original time series in order to optimize the accuracy of the represented time series [11].

- **Prediction** Prediction is one of the most commonly used tasks in TSDM and one of main area in data science fields. There are great demands in real world cases such as market behavior forecasting according to commercial data. Time series prediction targets to pattern feature correlations in a model and predict the future values of events.

  There have been many approaches and algorithms for generating predictions of future values of time series. Mostly they are focused on producing a single predicted value. For example, time series prediction has been approached using Auto Regressive (AR), Auto Regressive Moving Average (ARMA), Integrated Moving Average (IMA), and Auto Regressive Integrated Moving Average (ARIMA) models and neural networks (NNs). All of these approaches predict the next individual value in the time series. The prediction techniques which are mostly common used are ARMA model and especially Seasonal Auto Regressive Integrated Moving Average" (SARIMA) model [12, 13].

- **Classification**

  Classification is finding a function or building a model. It is an ability of assigning data to one class from many predefined classes based on their features. The model can be represented via mathematical formulae, classification rules, decision trees or neural networks [14].

  Classifications are used in different applications like pattern recognition, spam filtering, medical diagnosis,image and detecting malfunctions.

  This study focuses in time series classification because classification is likely to be the most known and common data mining approaches in all-time series mining tasks and it is assigned to supervised learning algorithms in machine learning.

2.4 Time Series Classification

The idea of time series classification looks like human understanding of similarity. Time series classification explains how to predict class label from the unlabeled

time series data using trained model from known data samples and known class labels. Therefore, the algorithm should be trained in advance by using the samples of the labeled data. The aim is to learn the special features which differentiate each class from another. After entering the unlabeled data into the model, the model will automatically recognize the class which belongs to the features [8, 10].

Over the past decades, the classification of time series has been growing interest, and arising in many fields including machine learning, data mining, statistics, signal processing, environmental sciences, economics, computational biology, image processing and chemo metrics. Many researchers focused on time series classification. According to most of the classification approaches, firstly, training model should be build based on labeled data, then the resulted model will be used in predicting and identifying the class labels of the unlabeled time series data [15]. Geurts clarifies the time series classification based on finding the local properties and patterns and then combining these features to build the model [16].

2.5 Time Series Classification Algorithms

Two most common approaches for time series classification are: instance-based and feature-based approaches. Instance-based approaches utilize the similarity between time series. While, feature-based approaches require extraction of meaningful information from the time series to be used as an entry to supervised learning algorithms.

2.6 Time-Series Similarity Measures

Instance-based approaches measure the similarity between two time series data sets. Similarity measures compute the distance between two time series which an indicator of level of (dis)similarity . They are the backbone of time series data mining as they are required nearly in all data mining tasks.

Various concepts for time series similarity measures have been offered in data mining by defining by defining and computing a distance function, dist(M, N), between two given time series M and N which plays essential role in terms of classification accuracy.

The two majors time series distance measures are Euclidean Distance (ED) and Dynamic Time Warping (DTW) which normlay used in classification. DTW is similar to ED distance with an extension as it propose nonlinear time scaling which called warping [17].

2.6.1 Euclidean and Dynamic Time Warping Distance

NN classifier with ED is used in feature distance space to compare a feature vector distance against other time series in the dataset. Euclidean distance is a fast metric which modeled on comparing and computing the a difference (distance) between original values in the ith point of the first time series and the actual values in the ith point of the second time series which is a one-to-one mapping of the two sequences. But it is not directly designed to detect pattern variations.

In general, ED has some draw backs especially when the data is time shifted so that ED does not produce precise results. It is known of its weakness in sensitivity to distortion in time axis, and as our dataset is function of time, Euclidean distance does not work well. So the DTW distance measure came as a solution to this special weakness of Euclidean distance measure. This similarity measure flexibility allows a one-to-many alignment in a non-linear manner. Similarity can be measured between two time series in which each one may be different in speed by the measuring the optimal warping alignment of each time point pairs. DTW is perfect towards time series models which shifted in time or distorted in size/shape. To align two sequences of X and Y coordinates using DTW, an $m \times m$ matrix is built where $(i_{th}, \ j_{th})$ element in the matrix is the ED $d(q_i, c_j)$ between the two subjects q_i and c_j [18].

2.7 Nearest Neighbor Classification

Background of K-Nearest Neighbor (KNN) method

The KNN algorithm is a traditional and simplest discriminative classification method for non-parametric pattern as it directly models the decision function which classify problems according to the similarity measure. KNN is also known with its dependency on distance function that required for data classification. Fundamentally, KNN classify an object using some related features which describe its class. That

according to simple majority vote of the nearest neighbors in each class frequency among k nearest neighbors in the classified data set.

Searching for similarity between time series is a simple task for time series classification. For example, KNN uses a distance function, "$dist(a, b)$, between two time series a and b to find the k most similar training observations $a_1, a_2, ...a_k$ to a query instance b"[19]. The class mode among the k most similar instances is then predicted to q. The signature is identified by a majority vote of its neighbors with k nearest neighbors, thus it is an instance-based learning [20]. The 1-nearest neighbor (1NN) classifier is categorized as an accurate method compared to many approaches that can be applied in time series classification. In this study, it is focused on 1NN classifier in the time series data by building 1NN model on full length of time series data in order to make accurate prediction.

Researches showed that nearest neighbor using DTW, which is a distance measure of KNN classifier, achieves competitive classification accuracy results. This has been found very efficient and successful on time series classification although the computing speed is heavily affected by the associated DTW algorithm.

2.8 Support Vector Machines

Support Vector Machines (SVMs) is a very popular, discriminative, successful and effective classifier with wide applications in machine learning that include: bioinformatics, visual machine, time series analysis and text categorization . The strength of SVMs is built on the base of finding the maximum margin decision boundary (hyperplane) among class regions. SVMs works in the high dimensional feature space and learns the classification task in that space using a kernel function without any additional computational complexity. SVMs is one of the kernel-based feature identification methods [21]. Cortes and Vapnik put the basis of SVMs according to statistical learning theory (SLT). Some research studies found that SVMs does not perform well in time series because SVMs fails to build optimal decision boundaries. (SVMs) shared in solving the diversity of learning and function estimation problems [22]. Special properties of SVM are instantaneously minimizing the experimental classification error, maximizing maximum margin classifier, kernel representation and margin optimization.

2.8.0.1 Tree-Based Approaches

In general, in a classification tree, each observation is predicted which belongs to the most repeated class of training observations in the region. The most important and interested results in class prediction in classification trees is the class proportions among training data in the terminal node region and the class corresponding to a special node region.

A random forest (RF) algorithm is created firstly by Breiman. RF is defined as building an ensemble of multiple predefined decision trees grown independently and in parallel. For each tree from a bootstrapped sample which is being trained on the learning data, class label is predicted by the most common occurring of the most trees. Trees grouping perform well when individual trees are uncorrelated from others. Dissimilarity is obtained among single trees using two sources for randomness. Firstly, each tree is modeled using independent training samples with replacement. That combines the concepts of bagging approach, which results in smooth prediction and reduce prediction variance without sacrificing the accuracy. Secondly, at each node in the training tree, only a random subset of data features are chosen. Ensemble methods like random forests have shown better performance in terms of accuracy than individual decision trees in classification and regression tasks [23].

Random forests are capable of detecting correlation between predictors. Moreover, random forests can provides information about the importance of features. Predictive variables (features) may be numerical or categorical. In our studied dataset, as all predictive variables are numerical.

The random forest technique estimates the significance of an expected variable by the increasing amount of the OOB (out of bag error rate) error when OOB data of that variable are replaced and keeping all other variables the same. The increase in OOB error is dependent on to the expected variable importance.

Random forest classification model consists of multiple trees. If the number of trees increases, the classification accuracy of random forest ensemble model will be increased till a specific (optimal) number of trees. After that, increasing the number of trees is useless and not significant. The optimum number of trees can be selected through trial and error experiments. That can be implemented in the

R package in simplest way. RF becomes increasingly common in wide fields of applications of machine learning like chemical engineering, biological science, agro science, medical analysis, finance, etc, and it also shows competitive prediction performance [22].

2.9 Feature Extraction and Selection

Feature extraction and selection is a main task in the classification or clustering process. Feature extraction is a process of transforming the original features to produce new relevant features using statistical formulas. On the other hand, feature selection is a process of detecting the most valuable and efficient sample of the input features to use in classification/clustering. The objective of those techniques is to obtain a suitable subset of features for classification or clustering usage.

2.9.1 Importance of Feature Extraction

Feature extraction is a vital process in classification technique. The important characteristics of feature extraction are summarized as follows :

– **Minimizing the cost**

In the true world, there are many applications that deal with too high dimensional data and that could be expensive in terms of processing and storage costs. In addition, in huge sample size, the measurement of features is overpriced. One research found that feature discovering has a vital role in reducing the cost of features extraction and that will maintain excellent classification results.

– **Data visualization**

The main concern of feature extraction is saving the distance information and arranging the first main data in two or three dimensions. For descriptive objectives, it is helpful and suitable to graph the high dimensional data to more than two dimensions. The normal method in data visualization is linear projection.

– **Dimensionality reduction**

One of the main causes for the problem of dimensionality is that high dimensional computation have possible calculations and storage problems than low dimensional ones. Another reason is the noise produced from high dimension data which cover the the real patterns of models, which makes the classification approach harder. So those problems need to be taken into consideration in time series data mining as one instance of a time series is observed as one dimension, so that dimensionality of huge time series is typically super high.

Classification is considered to be more precise if there are many features. Many features means having good information based on having infinite number of samples. However, with enough samples, some of the dimensionality problems can be tackled.

2.10 Time Series Data Mining Applications

Time series mining is a vital data mining task for many real-world applications. Time series databases can be seen in almost every applications such as weather forecasting, science experiment, stock market research, medical diagnostics, environmental monitoring, manufacturing and production, temperatures in data centers, and physiological signals in health care.

Many researchers handle different applications for time series mining for example sensor measurements, mobile tracking, eye-hand tracking, data center monitoring, motion capture sequences, climate forecasting, environmental monitoring (like chlorine levels in drinking water), banking(loan/credit card approval) to predict good customers based on old customers. In addition, fraud detection(telecommunications, financial transactions) by identifying fraudulent events from an online stream of event, economic and financial time series where the user may be interested in event patterns, for example, the user might be interested in the preceding events of a large market crash and musical querying to detect if there is coping of the original music, and many more applications [10].

The value of data mining techniques are rising up since new developments and technology are coming to light, which allow getting huge massive of time series

data with trillion observations and more. This research is challenge to deal with such large time series with proposed approaches.

Chapter 3

Literature Survey

Biometrics research is rapidly increasing because of the demand of biometric applications in user computer interface, security, and other related areas. It is agreed that biometric recognition systems require collective analysis of various behavioral or physiological features. Those are considered to be the most effective and flexible systems in individuals authentication. Thus, in that case there is no need for password reminders or carrying smart cards. The researches mainly deal with different biometric applications like iris, retinal, finger prints, hand writing gesture, speech identification signature, voice and other biometric characteristics. Those are characterized by uniqueness which is useful and can be used for authentication processes and identification.

Since the last decade, mouse dynamics have attracted more and more researchers; some studied the movements and gestures of the mouse. They tried to extract available features that help in user identification [14, 24, 25]. These researches are a little bit similar to our study while other studies interested in the coordination of mouse and eye movements [26, 27].

In the new era of security, securing mobile devices has considered as a big challenge to identify the users [28]. Identification mobile users using interesting biometric patterns has been are attracted many researchers. Abundant researches have proposed many authentication mobiles applications. The authors in [28] proposed classification model for mobile user identification by capturing user tasks, using mobile apps logs and combining user behavioral features. They consider four basic elements in their proposed model regarding to how, where, when and what,

those refer to gesture or input-output, location, time and apps usage. That study is in some way different from our experiment which captures the movements of users gaze and mouse to build identification model according to those movements. While Song [29] presented EyeVeri "eye movement based authentication system " for securing smart phones. It captures human eye movements through frontal built camera and then they evaluated the system by employing SVMs as a classifier. That algorithm showed high accuracy for authentication process [29], whereas random forest model is preferred in our experiment.

On the other hand, the intention of Aksari and Artuner [30] on their research experiment at 2016 is similar to our aim. The goal is to replace password usage authentication through application for users authentication using mouse movements. They extracted the mouse features from paths between user clicks, which are shown on a screen. Then, they built feature vectors by doing computational processes such as normalization, finding the speed of the mouse, the acceleration of the mouse and the angle of the mouse movement. By using nearest neighbor algorithm, they got success rate around 92% [2, 30, 31], which considered as high accuracy.

On the other hand, in [32], they studied mouse behavior and proposed new characteristics to obtain the feature vectors. Then, they used SVMs for classification and authentication process, and the experimental classification accuracy result was 96.3%. The approach in those studies can be applied in real-world dynamic soft keyboard scenarios such as logging into online banks and instant messengers by using soft keyboard even if the user password has been stolen [31]. In [32], the authors discussed problems about security issues and passwords attacks. They proposed methods in order to enhance the security against these attacks by using handwritten signatures of mouse movements for authentication. Mouse signature digital verification and authentication are approached through mouse actions capturing and recording, feature extraction and classification by applying neural network concepts on the feature vectors and building user profile model. On the other hand, [33] proposed another solution for traditional password problems using graphical passwords authentication. They conducted eye tracking experiment of the -Image Pass- concept. In addition, Weaver and Mock [34] proposed Eye-Dent system. Users can enter passwords by looking at each symbol in the password consecutively without cutting in screen keyboard. After that the gaze positions are gathered to find the users chosen numbers.

Another research focused on solving fraud authentication problems and biometric attacks [35] with new and challenging method. They discovered the uniqueness of nail biometric because the finger nail bed shows great percent of individuality in all cases like identical twins and even in the same hand between different finger nails of a person. Since the features of the nail are unique for every subject, it is used for authentication and identification. The authors experimented this by using classification algorithms such as SVMs classifier and Naive Bayes, and they presented high accuracy for both.

A research study in [36] explained the usage of behavioral sensors in computer systems for active authentication through detecting the features from mouse and keyboard actions. Then, they developed classification model using Naive Bayes classifiers.

A major challenge in [37] is to address cheating issues in computer games which occur for gaining money from selling virtual assets or hackers who steal money from stolen accounts. In his research, he studied the possibility of identifying two computer game players based on their mouse and keyboard dynamics by extracting their features and then applying nearest neighbor classification method for identification process.

Youming's [38] aim is to use saccadic eye movements authentication for personal devices like a computer or mobile phones instead of password since saccadic eye movements is hard to imitate, and easy and fast to compute. He used many classification methods for performance verification to recognize an authenticated user from the extracted features such as kNN classifier, neural networks and SVMs, and high accuracy was obtained, which was close to 95%.

Recently, Kasprowski and Harezlak [39] combined the eye with mouse movements for behavioral biometrics. They aimed to find a good solution for user identification according to eye and mouse movements. They extracted the important features for the input data, built the DTW distance matrix, applied SVMs algorithm for identification model, and they got high accuracy for eye and mouse combination .

Finally, it can be summarized that the need for secure human interface becomes an urgent requirement in this age of high technology. Therefore, this study tries to find a perfect solution as a substitute for traditional passwords in authentication process

using classification models such as RF and NN. In addition, this study shows that fusion between eye and mouse movements is good opportunity for building secure biometric authentication system with high accuracy.

Researches can be categorized according to behavioral and physiological features. As mentioned before, the researches which concern of identifying behavioral features are: [26–28, 33, 34, 36, 36, 37, 39] while the researches that interested in identifying physiological features are: [35, 38].

Recently, mouse dynamics attract a lot of researchers as it is simple biometric authentication method and cheap. So that they try to develop that approach by fusing it with other devices like keyboard or eye tracking devices which provide with more accurate authentication models.

The differences of those works is that some of them just applied the experiments into limited subjects and built the authentication model using SVMs or KNN without performing features extraction procedure. On the other hand some researchers used the statistics feature extraction method in order to build authentication model. Still, there are some differences in the work done as this study build distance matrices for new extracted features like acceleration and speed matrices. Also many features are extracted which fed in to RF model and NN model that produce good and competitive results.

Chapter 4

Data and Analysis

4.1   Data Description

When mouse and eye inputs are handled, usually the location of the mouse pointer, eye gaze directions and the state of the mouse buttons are required to be known. In this experiment, a log-in screen with group of numbers is displayed in front of user. Instead of entering a password the user is asked to use the mouse click by selecting numbers shown in the screen by tracking the mouse and eye movements toward the screen. In the selection process, the user will look at the numbers in the screen and he has to follow his gaze direction by mouse click to choose the numbers. When the mouse cursor stopped at a specific number, that will be recorded and then select the following number. The same procedures will be repeated for the other numbers with the gaze directions and mouse movements. So a sample of subject's mouse and eye movements has been collected in the log-in screen in that way. The aim is to discover the features which could provide a predictive link between where the subject is currently looking at the screen and where the mouse is positioned. This research focus on using mouse movements and gaze directions related to biometrics to find the link in order to improve existing recognition algorithms. In my research, mouse and eye movements are considered as biometric. So the mouse here, is not used for signing but for authentication and verification the subject. The traditional authentication by log-in password method is replaced by mouse movements authentication since each user has its own patterns. Those patterns are utilized for identification process for that user. The raw data sets consists of different number of classes distributed into three sessions

26

TABLE 4.1: Number of actions in the Used Data sets through Experiments

Name	Number of subjects	Mouse movements	Eye movements	Mouse click
Data A	27	18462	9282	324
Data B	30	65288	11979	360
Data C	24	15421	8183	288

for each class subject with three features belongs to all classes as explained in (Table 4.1).

4.1.1   Data Variables (Input Features)

– Class subject

– $V1$: Measured time by milliseconds for the recorded movements (Table 4.2).

– $V2$: Type of recorded movements

  * M for mouse position

  * G for gaze position

  * MC for mouse click position : A mouse click happened by pressing the middle button or the scroll wheel of the mouse (there are exactly 4 mouse clicks in each trial).

– $V3$: X coordinates for mouse and eye movements

– $V4$: Y coordinates for mouse and eye movements

TABLE 4.2: Example of the raw data

V1 (Time)	V2(type of movement	V3 (X coordinates)	V4 (Y coordinates)
0	MC	179	128
53	G	45.3	-17.8
118	G	454.1	67.8
120	M	180	130
128	M	193	131
135	M	206	133
144	M	221	136
151	G	477.4	55.3
154	M	236	137
160	M	251	140
167	M	269	142
176	M	292	146
184	M	317	146
185	G	482.0	59.1

4.1.2   Datasets and the System Used in Experiments

The comparative experiments have been performed for a three different time series datasets which obtained from look and click competition website and divided into three different sessions [40]. All of the datasets in that competition are publicly available and labelled. Our data has various features and determined number of classes. The used approaches are coded and implemented in Intel (R) Core TM i5 2.40 GHz PC with 4.00 GB RAM using R studio-1.0.136 -64-bit. On other hand another PC Intel (R) Core TM i7 3.60 GHz PC with 32 GB RAM is used in order to build the 1NN model based on DTW distance matrices which needs huge memory and run time -around 12 hours- per each matrix.

4.2   Data Interpolation

The studied data sets are non-uniform and unevenly spaced time series. So it is required to transform the data into equally spaced observations using linear interpolation. Linear interpolation is a method estimating the missing values between known data points. The codes are explained in more detail below :

```
# Transforming irregular time series to regular time series
Algorithm: Zoo package
Input : A data set X
 Make List for X
 for For i = 1, , , , m   #m is the number of observations.
read   tables inside the list .
 Order   and filter the data subjects   according   to   mouse movements.
 Order the mouse data subjects according to its X coordinates and time
    .
 Build regular time series from min time to the maximum.
Estimate   the missing values
 end for
Output :   Mouse X coordinates time series
```

LISTING 4.1: The pseudocode of transforming irregular time series to regular
time series

4.3   Classification

Classification approach is the best approach for analyzing the time series data since many class subjects are available and the need is assigning the subject label for each time series. On the other hand allocating the important features to identify subjects from others which helps in building good model. There are many classification algorithms that can be used for generating classification models.

4.3.1   Similarity Measures

The similarity measures are usually used to compute the similarity between two time series and find the accuracy of the time series model.

Euclidean distance is a difference between two actual values in two different time series, for example finding distance between i-th point of first time series and the i-th point of second time series.

In the feature distance space, the nearest neighbor classifier with Euclidean distance is used to compare distance feature vector against those in the dataset. The Euclidean distance is a one-to-one mapping of the two sequences. Generally, Euclidean distance do good job but it is still does not always give in high accuracy output especially if the time series data is a little bit moved by time. It is sensitive and weak to deformation in time axis. Our dataset is function with time so Euclidean distance did not work well, therefore DTW distance measure came to solve the special weakness of Euclidean distance measure. The flexibility of this method allows a one-to-many alignment in a non-linear manner and measure the distance and homogeneity between two time series where each time series has different speed . The total distance between those time series could be minimized by finding path through them. To align two sequences of X and Y coordinates using DTW, a $n \times n$ matrix is built with the (i-th, j-th) element of the matrix. [13, 41].

4.3.2   Nearest Neighbor

Nearest neighbor as explained in previous chapters, depends on distance of data observations for classification algorithm. Mainly, the KNN classifier identify the class subject using a set of relevant features by the simple majority vote of the nearest neighbors of that class among KNN in the classifiers dataset.[23]

The proposed work is split into two modules such as:

- Building up the DTW distance matrix based on X-Y coordinates and time for mouse and eye. They are explained in more details in the following section.

- Finding the similar subjects from the distance matrix constructed.

4.3.2.1   Data preprocessing

1. Three input features for each data set are available and each data has three sessions. Then, three subjects lists are prepared, list for X coordinates , list for Y coordinates and list for time for all mouse and eye movements.

2. **Standardization**, scaling all features in the data set and transforming all the values into a common scale and same range. This is required for data preparation since with normalization problems of different scales can be avoided.

The scaling method is indication to standardization or normalization and can be computed using the formula

$$Z = \frac{X - \mu}{\sigma} \qquad (4.1)$$

Normalization formula is:

$$Z = \frac{X - \min(x)}{\max(X) - \min(X)} \qquad (4.2)$$

where Z is the normalized value. X is the old value, μ is arithmetic mean of variable vector $N(\mu, \sigma A)$, and σ is the sample standard deviation of the variable vector and after normalization $N(\mu, \sigma A)$, the random variable will follow normal distribution with mean and standard deviation of $N(0, 1)$..

After, normalizing the data nearest neighbor distance matrix is built on the normalized data input features (X coordinates, Y coordinates and Time) by computing the distances between every pair of subjects in the dataset and stores them in an $n \times n$ matrix. This distance matrix is created for classifying subjects to speed up the process during the classification task . Still, it was noticed that matrix codes run time is extremely slow as it took approximately 12 hours for building a matrix. Dataset A is taken as an example for clarification which has 27 subjects distributed into three sessions. So 81×81 matrix is build.

The output is an $N \times N$ matrix where each $i; j$ entries contain the distance between points x_i and x_j for the all set of N in the data set.

After that, the minimum distance is chosen between each two subjects in rows and columns in the constructed matrix.

4.4   Feature extraction

4.4.1   Introduction

Feature extraction is the core of identification system and the operation of converting the original features to new ones which determine certain characteristics of objects [42]. Features are useful for identifying class subject from another. Feature space is always with high dimension. So that the goal of feature extraction is trying to reduce the high dimensions of the feature space into suitable space for classification methods. The features should be effective and capture the essence of what humans consider in similar sequences. That computation should be fast and scale-able[43].

To reduce the high dimension of data, feature selection is performed for random forest model which is a very important process of detecting the most important and effective input features that will be used in classification and recognition model. Also it is able to extract more information while reducing noise and avoiding redundant data with fast computation. By taking the raw data and applying calculations to extract characteristics that signifies subject behavior. A feature matrix is created that can be used to gain statistical information using supervised and parametric learning techniques.

4.4.2   Constructing the Dataset (Data Processing)

Based on the collected raw data in Table 4.2, several features for classification have been created which help in identifying subjects uniquely and could provide a predictive link between where the user is currently looking at the screen and where the mouse is positioned. Mouse dataset contains data based on cursor movements, double clicks, speed, clicks and click locations, scroll wheel usage, directions, etc. The features that have been extracted from the original data are statistically analyzed by computing three values for each class because three sessions of data are available. Features examples: sum, the minimum, maximum, mean, standard deviation, difference, speed, acceleration, range and others operations are computed. This analysis produces seventy four features for all mouse and eye movements and from the first glance, it seems that all available features should be used to recognize users, still, using all features may worsen the accuracy of the model as some of these

features could be redundant or irrelevant to have perfect classification model. So, in order to produce such model, the most important and related features should be selected.

The complete lists of the extracted features are provided in Table A.1.

The main data features are X coordinates, Y coordinates and Time (T) and according to that, other features are statistically extracted. So the new features could be classified as:

1. Features based on X coordinates

   – Scale_x: The mean of normalized X coordinates position for all subjects.

$$Z = \frac{X - \mu}{\sigma} \tag{4.3}$$

   – SD_X: The mean of standard deviation for X coordinates position for all classes.

   – Mean_diff_x: The mean of the difference between two consecutive points of X coordinate for all classes.

   – Max_x: The maximum value for X coordinates position for each subject.

   – Min_x: The minimum value for X coordinates position for each subject.

   – Mean_x : The average for X coordinates for each subject.

   – Sum_x : The summation for all X coordinates positions for each subject

2. Features based on X Y coordinates

   – Length : The sum of the distances between all close curve coordinates for each subject. Mouse curve (c) has a length of the total moved distance with n points and it is represented by :

$$Length(c) = \sum_{i=1}^{n} \sqrt{(X_i - X_{i-1})^2 + (Y_i - Y_{i-1})^2} \tag{4.4}$$

   – Speed : The proportion of the total distance that the mouse traveled from one point to another divided by the total time taken to complete the movement and can be represented with the next formula:

$$Speed(c) = \frac{1}{n} \sum_{i=1}^{n} \frac{\sqrt{(X_i - X_{i-1})^2 + (Y_i - Y_{i-1})^2}}{(T_i - T_{i-1})} \tag{4.5}$$

– Acceleration: The acceleration of the mouse movements which computed by finding the speed divided by the total time difference between continuous two points and represented by:

$$Acceleration = \frac{\Delta v}{\Delta t}$$

3. Features based on Y coordinates

– Scale_Y: The mean of normalized Y coordinates position for all subjects.

$$Z = \frac{Y - \mu}{\sigma} \tag{4.7}$$

– SD_T: The mean of standard deviation for Y coordinates position for all classes.

– Mean_diff_Y: The mean of the difference between two consecutive points of Y coordinate for all classes.

– Max_Y: The maximum value for Y coordinates position for each subject.

– Min_Y: The minimum value for Y coordinates position for each subject.

– Mean_Y : The average for Y coordinates for each subject.

– Sum_Y : The summation for all Y coordinates positions for each subject.

4. Features based on Time

– Scale_t: The mean of mouse normalized time for all subjects.

$$Z = \frac{T - \mu}{\sigma} \tag{4.8}$$

– SD_T: The mean of standard deviation for time mouse/eye movements for all subjects.

– Mean_diff_T: The mean of the difference time between two consecutive points for all classes.

– Max_T: The maximum value for value for time mouse/eye movements for each subject.

– Min_T: The minimum value for time mouse/eye movements for each subject.

– Mean_T : The average time mouse/eye movements for each subject.

– Sum_T : The summation for all time movements for each subject.

The other extracted features are related to eye movements and mouse-click actions, which will be similar to mouse features as they are based on same statistical functions.

4.5 Data Visualization

4.5.1 Mouse Positions

The different records for X and Y coordinates mouse movement for the three sessions are shown in Figure 4.1 for all subjects, which means that mouse behaviors for the same user could be change on different times (sessions). This indicates that mouse movement not fixed for the same user with different time sessions. The subject change his behavior in the way of using the mouse, which mean on different time sessions the subject may be choose different passes in the screen in order to select the same numbers.
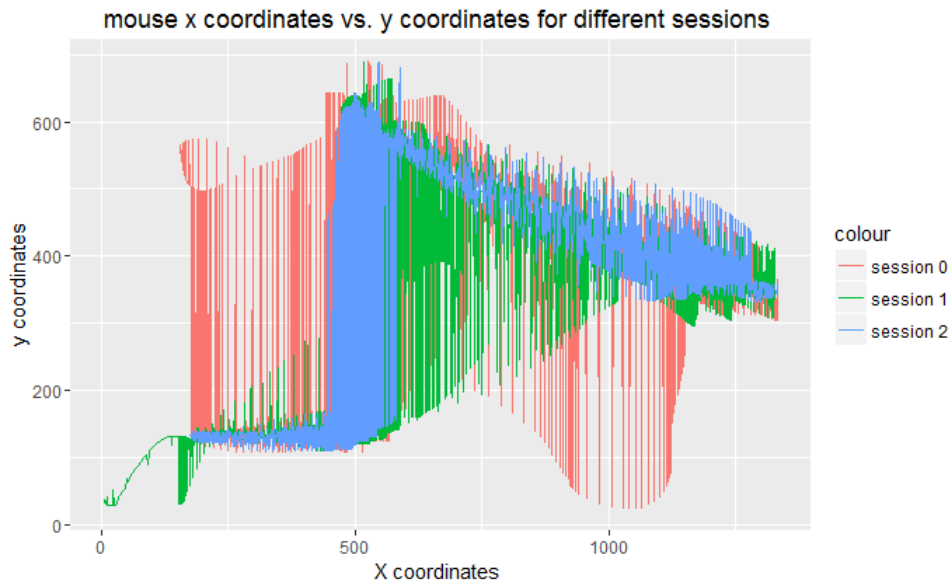
FIGURE 4.1: The relation between mouse X coordinates and Y coordinates for different sessions

4.5.2 Gaze Positions

Figure 4.2 below shows the gaze positions according to X and Y coordinates and the different colors refers to the different times sessions, which indicate that user behavior of eye movements change with different times. The eye positions occasionally follow the mouse cursor in the screen. On the other hand, the subjects change their eye gaze positions in the selection process on different time.
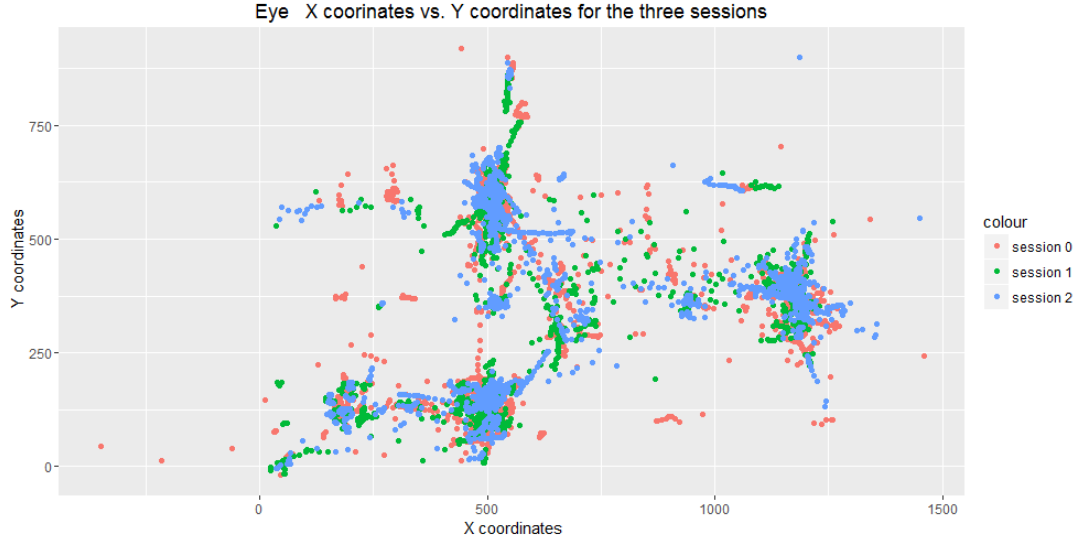
FIGURE 4.2: Gaze positions on X-Y Axis

4.5.3 Normalized X Coordinates

Figure 4.3 shows the normalized X coordinates for mouse and eye movements. There are gaze X coordinates positions related to some subjects are upper than their mouse coordinates positions. Other subjects have gaze coordinates are lower than their mouse positions, which indicate that there is no big correlation between mouse and eye movement. Some users look at some points on the screen while the mouse cursor is far from these points.

FIGURE 4.3: Normalized X coordinates for mouse and eye movements for all subjects

4.5.4 Normalized Y Coordinates

Figure 4.4 shows the normalized Y coordinates for mouse and eye movements. From looking to the figure, it could be summarized as there is no correlation between the gaze direction and mouse direction. Still, the gaze direction and mouse cursor at some points are close to each other but that does not provide any additional value or useful information.

FIGURE 4.4: Normalized Y coordinates for mouse and eye movements for all subjects

4.5.5 Length of the Curve

Figure 4.5 below shows the different length curves for each subject according to their eyes, mouse and mouse click. The distance between cursor and gaze positions

is long at some points for all subjects while at other points it is generally shorter when the cursor is placed over the selected number. This could be summarized that when the user want to select a number from the screen by clicking the mouse, the gaze direction will be near to that target number and before the mouse cursor, otherwise the mouse movements curve is almost shorter than eye movements curve.
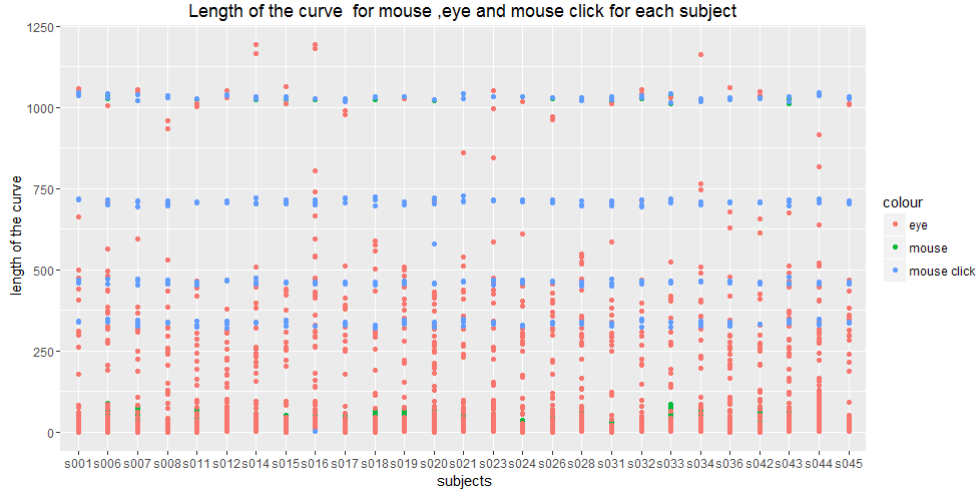
FIGURE 4.5: length of the curve for mouse and eye movements for all subjects

4.5.6 Speed of Mouse and Eye Movements

Figure 4.6 displays the speed values for mouse, eye and mouse click for each subject. Most of times the speed of the eye is the higher than mouse cursor movements. Moreover it is also faster when the subject want to select target number by eye gaze than with the mouse movements. The interesting to see is that the mouse spends a much higher percentage of time in some regions than the eye, which means that eye gaze selection showed some slowing as in figure (between zero to 10 msec.). Apparently, at some moments when the subject clicks the mouse to the select a number, almost the speed of mouse click and eye are near to each other even though the eye still lead the mouse. In conclusion generally, subjects eye gaze movements technique is faster than mouse movements. In addition our subjects were comfortable in choosing the target numbers from the screen with their eyes, but there was some slight slowing of performance with eye gaze that might indicate some stress or fatigue.
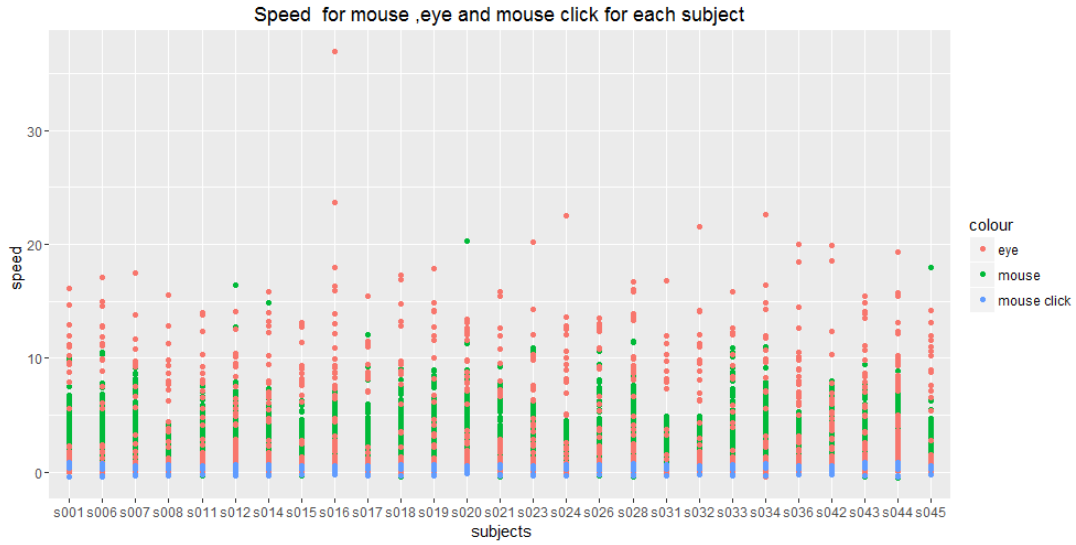
FIGURE 4.6: Speed for mouse and eye movements

4.5.7   Acceleration

It is known from previous Figure 4.6 that eye can move faster than the mouse. So when the mouse being moved faster by the subjects by accelerating the speed of mouse with respect to time, the results for this acceleration will be as in Figure 4.7. Obviously, when the user apply more force to the mouse to make it faster that leads to acceleration for the mouse is faster than eye. So if they slow down the speed as they get closer to target number as shown in the bottom of the figure, the acceleration is smaller and the mouse cursor near to the gaze direction.
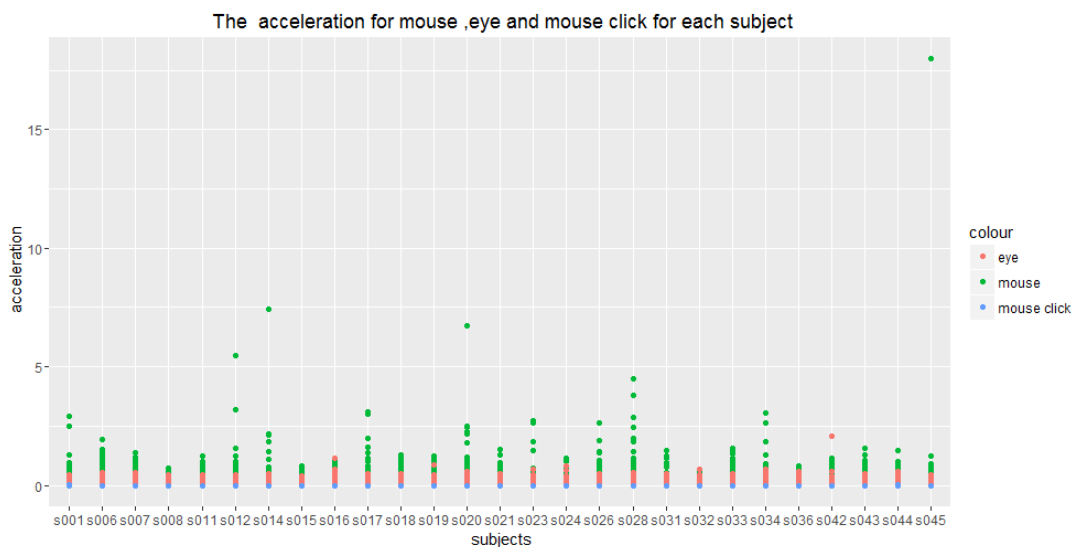
FIGURE 4.7: Acceleration for mouse and eye movements for all subjects

4.5.8 Average Time

Figure 4.8 shows that the subjects require more time for selecting the target number by their eye than by their mouse. That indicates slowing of performance with eye gaze which might indicate fatigue. The time can be varied across different mouse events for example mouse click time in some subjects is higher than the mouse cursor movements time and in other subjects the mouse movements take more time but in general case the mouse click time is almost near the average time for eye movements.
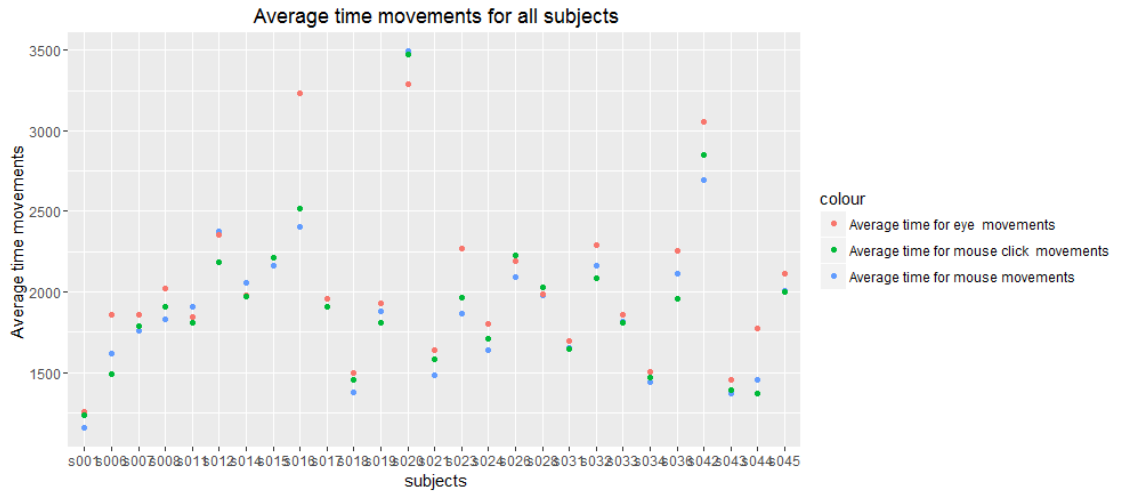
FIGURE 4.8: Average time for mouse and eye movements for all subjects

4.5.9 Euclidean Distance

Euclidean distances is considered from the simplest time series similarity measures. The difference between X coordinates for each eye and mouse position is Δx and Δy is the distance between Y coordinates for all eye-gaze and mouse position.

For each point of time the distance between eye and mouse coordinates is plotted in 4.9, which shows that there is relation between gaze and muse positions. It shows the different values of euclidean distances for each subject for mouse and gaze coordinates.

On other hand at some points it could be seen that their Euclidean distance are near zero which indicate that some subjects cursor positions and gaze positions are so near which approximately zero. Also the points which larger than zero means that the mouse cursor is upper the gaze position on the screen as the user

at sometimes does not look at the same point where the cursor move and the gaze position will never be on the same point of mouse cursor, there always some limits they can be near but the distance will not be zero.
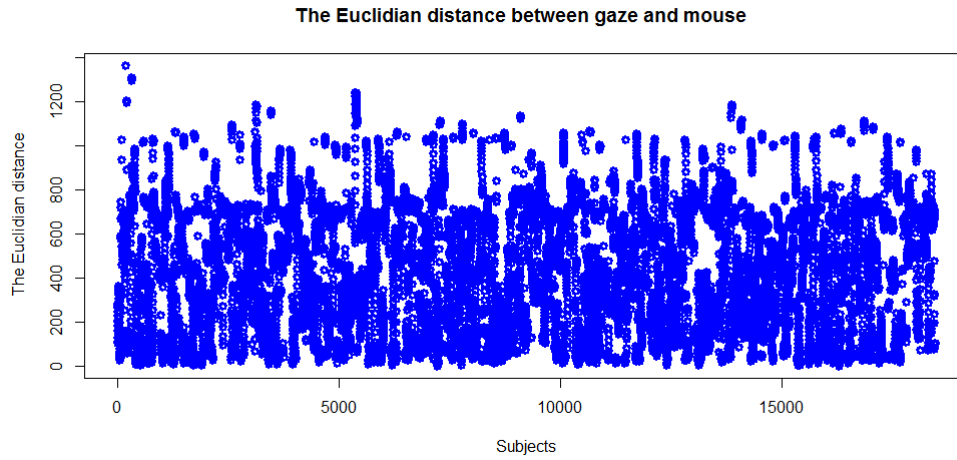
FIGURE 4.9: Euclidean distance between mouse and eye movements

4.5.10 Difference X Coordinates Y Coordinates for Eye and Mouse Movements (Δx,Δy)
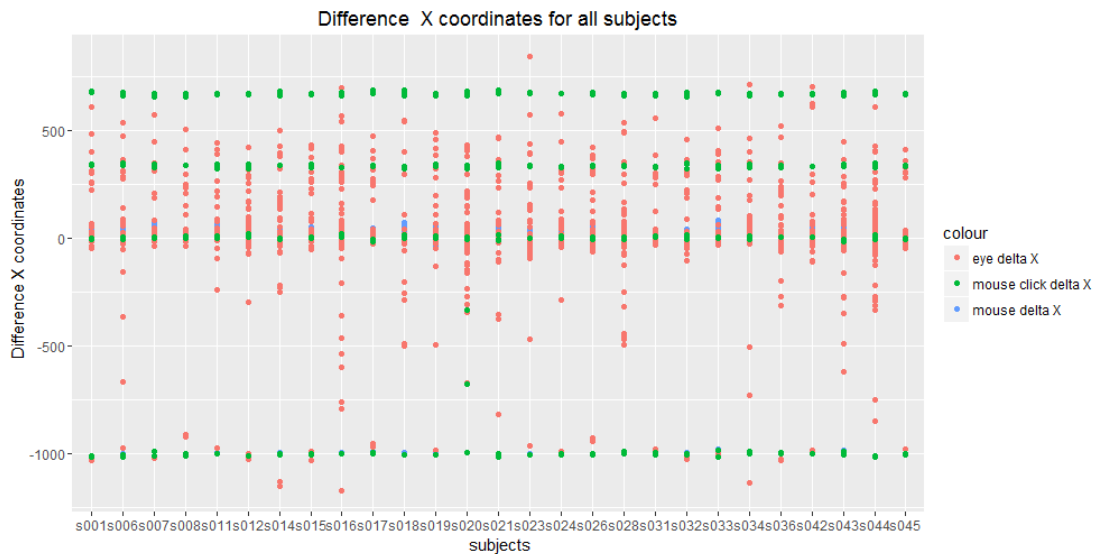
FIGURE 4.10: X coordinates difference for mouse and eye movements for all subjects

Figure 4.10 illustrates the difference of x coordinates for different movements. For more illustrating about Δx that users stop the cursor before or after their gaze in order to prevent the visual analysis confusion. As it is shown the mouse click Δx

is almost near the eye difference x coordinates for most of times. Subjects were often keep the cursor slightly offset from where the target number is, which is not close to particular part of the screen.
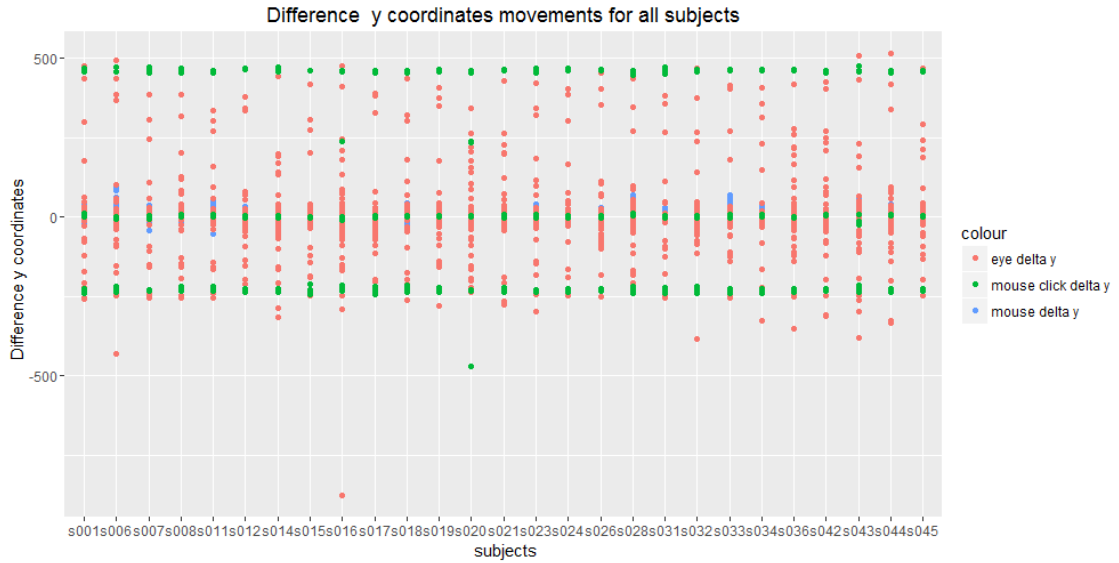
FIGURE 4.11: Y coordinates difference for mouse and eye movements for all subjects

The same explanation as Δx, Δy is the y coordinates difference positions for each subject where users put the cursor upper or lower their gaze position to prevent it from visual analysis confusion or partially hiding the target number. The different mouse y coordinates its very small its close to zero in most subjects which mean that at different times the subject keep moving the mouse within the same direction on the y coordinates 4.11.

Chapter 5

Experimental results

This chapter provides experimental results to illustrate proposed approaches.

5.1 Nearest Neighbor Classification Results

5.1.1 Measurements and Performance Metrics

Many DTW distance matrices are constructed according to the original features which are X and Y coordinates. Moreover speed and acceleration matrices are build based on the X and Y matrices. Multiple experiments using KNN are performed. DTW with 1-NN approach does not need train or test set for parameters optimization . Table 5.1 shows the accuracy of the nearest neighbor for different extracted features matrices such as the x coordinates, Y coordinates, velocity and acceleration. The classification accuracy is usually measured by computing the percentage of corrected subjects numbers. The extracted features were normalized and compared with nearest neighbor algorithm, which is explained in the last chapter. Different speeds and accelerations features are tested in R. Using X coordinates, y coordinates and time matrices, the error rates were non considerable because its so low for all data sets (Table 5.1). Even after trying another extracted features and implementing the nearest neighbor for them like velocity and acceleration, the performance of those two vector features were frustrated as the accuracy of the model near zero. So the need to try another competitive approach to get good results. The NN with DTW parameters, accuracy of the models and time spent at

the DTW- 1NN method are figured in Table 5.1. In addition, confusion matrices are listed in A.

The Analysis with k-Nearest Neighbor classifier yielded the following results:

TABLE 5.1: The classification accuracy for nearest neighbor

Name	Mouse -X coordinates accuracy	Mouse -Y coordinates accuracy	mouse Velocity	acceleration	Eye -X coordinates accuracy	Eye -Y coordinates accuracy	Eye Velocity	eye acceleration
Data A	0.0370	0.0123	0.0247	0.0123	0	0.0370	0.0123	0
Data B	0.0333	0.044	0.0333	0.0444	0.0111	0	0.0111	0
Data C	0.0417	0.0139	0.0278	0.0278	0.0139	0	0	0.0139

5.2   Random forest

5.2.1   Data Partitioning

The imbalanced data sets that we have lead to problems in learning and identification process so that many methods has been developed for such data sets like: random forests and cost-based optimization [44].

So in practice an accurate model with high prediction is required. In that case of supervised learning, a computational model is trained to predict the subjects. In order to get the results from the determined data set, the data set is split into training data and testing data. The percentage of train data classes should be 2/3 from all data sessions and 1/3 of data observations will considered as test set as shown in following pseudocode. For example data set A has 81 observations distributed on three sessions so the train data includes 54 observations and test data contains 27 observations.

On other hand there are three sessions for each subject so stratified sampling is used in order to make balance in our model and to make sure that each subject has entered the model. Stratified sampling can be defined as: "samples from each cluster are selected with a uniform probability". There are more possible ways, how to choose the number of samples to be selected per cluster. Initially, the data is split into two groups. The first group is used to train the model, the second group to measure the error of the model that should be one achieves the most accurate model. For instance, data A has 81 observations, sampling is constructed by randomly selecting 2 samples from each class with total 54 observations to build the model and the remaining 27 samples to measure that models error. To get the

classification accuracy, the predicted labels should be compared with the actual class labels using test data and that called the confusion matrix.

5.2.2 Design and Performance Improvements for Random Forest

RF algorithm is multiple classes approach, which build from aggregating many trees together. For example if training data contains N observations, then samples with size N are taken from it. In order to grow user-friendly RF model, many various parameters need to be adjusted. The main variables (parameters) in the RF model are number of grown trees (ntree) and number of randomly chosen features at each node (mtry) as explained in the following pseudocode. In order to improve the the model accuracy, those parameters should be optimized.

```
1  ####################################################################
2  #fit random forest with samplesize
3
4  Input : Featured data D
5  call D
6  apply randomforest function for  data D
7  choose the nodesize =1
8  extract the confusion matrix
9  print the accuracy
10 ##########################################
```

LISTING 5.1: The pseudocode for random forest for Data A

5.2.3 Random Forest Features

For better understanding the relations between subjects and the authentication system, a model should be constructed to analyze the data and features. Typically, this model would be both descriptive and predictive in nature. In practice features matrix is produced which contains 74 features for all subjects in which each subject has three sessions. Since the model is depended on the accuracy of eye and a mouse movements, the need to be sure that the same features for all users are produced.

5.2.4 Random Forest for Data A

Our task is to correctly identify the class labels of each data set using the RF model that will be built in this analysis. R is used, a free available program on the Internet, to construct the random forest model. These features can not be picked whole because that will affect the model efficiency. R is used to find the most important features which provide the optimum accuracy. For each feature set described above, RF is trained with more than 500 trees for two sessions from the 27 subjects data sessions (Data set A) and tested on the remaining session of the subjects in a leave-one-out cross-validation approach. The most important attempt is to predict the correct subjects according to the extracted features for the available subjects in the model by considering their label as a name of the subject. So learning model give the predictions for each subject and the most accurate model which predict all subjects without any errors. After that, the model is tested by predicting the ID of the subject using eye and mouse movements in the test sample. Comparing the predicted class IDs with the true class IDs, a confusion matrix is created with 27×27 dimension in which the element in row i, column j is counting the number of times the subject with true ID i is correctly predicted, to all ID j using the random forest.

5.2.4.1 Constructing the Model

After using the featured data set A which contains 81 observations then choosing 25 random variables (mtry = 25) and letting the number of trees (ntree=1000) as in following codes. The classification performance for the random forest classifier is examined using the various feature sets. The highest average classification accuracy using training data was 59.2% on feature set which means that identification model identify around 59.2% from the subjects correctly and the out of bag error OOB estimate of error rate is 41.98% which mean that around 42% of data did not enter the RF model.

```
1  #The    final  result  for  random  forest
2  Input  : Featured  data D
3  call D
4  apply  randomforest  function  for   data D with  the  following  parameters:
5  1.choose  the  nodesize  =1
6  2.choose  the  samplesize
```

```
7  3.choose the test data
8  4.number of trees ntree=1000
9  extract the confusion matrix
10 print the accuracy
11 print test error
```

LISTING 5.2: The pseudocode for random forest with validation data A

5.3 Results

These are the results which obtained by performing the all experiments which have been done as it is shown in table 5.2. All data sets are checked for all class subjects to get the total classification accuracy which calculated by comparing actual target labels with predicted target class labels. The explained results can be observed from the confusion matrices in which each row represents the number of the expected classes and each column represents number of subjects in predicted classes as in table A.3.Confusion matrix A.3 shows how subjects are predicted in classification. There are 47 from 81 classes are correctly predicted, which are belong to their actual class subjects and 34 out of 81 are falsely predicted to belong to other classes. Similarly, the classification model some times fails to predict the class label correctly so the classification error for was 33% falsely predicted from the total number of sessions, and 67.6% falsely predicted, and the model predict some classes 100% correctly. The important variables are used to rank the importance of all variables in RF model based on data A as shown in Figure 5.1. The measure for that is the Mean Decrease Gini which based on Gini impurity index which help in defining classes. In Figure 5.1 the most important variables are acceleration and speed for the mouse click.

The important features in random forest model for dataset A
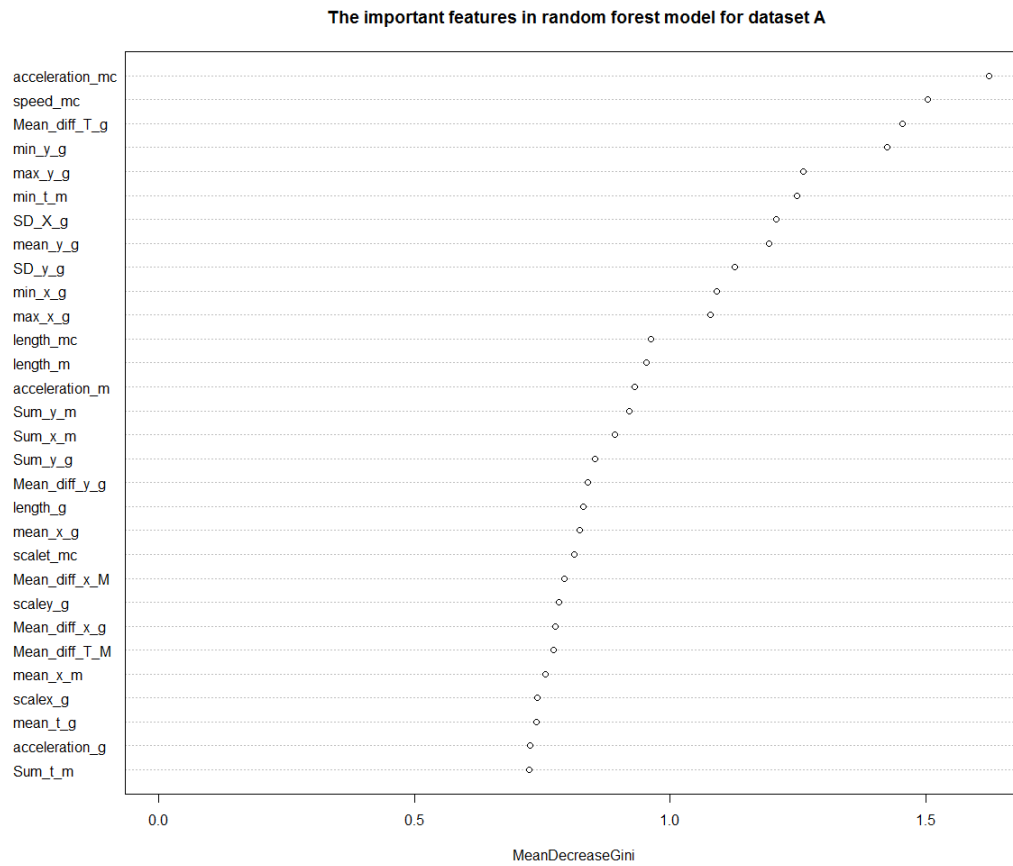
FIGURE 5.1: Important variables in RF model for data A

Figure 5.2 shows the importance ranking for RF variables which built on data set B. The most important variables are the maximum eye movements in X coordinates, the minimum eye movements on X coordinates and the minimum eye movements on y coordinates. After that, the distance of mouse movements and speed of mouse movement are important. Those are calculated based on Mean Decrease Gini, so the highest value of Mean Decrease Gini means that particular variable (maximum x coordinates of eye movements) plays a perfect role in splitting data according to their class labels .
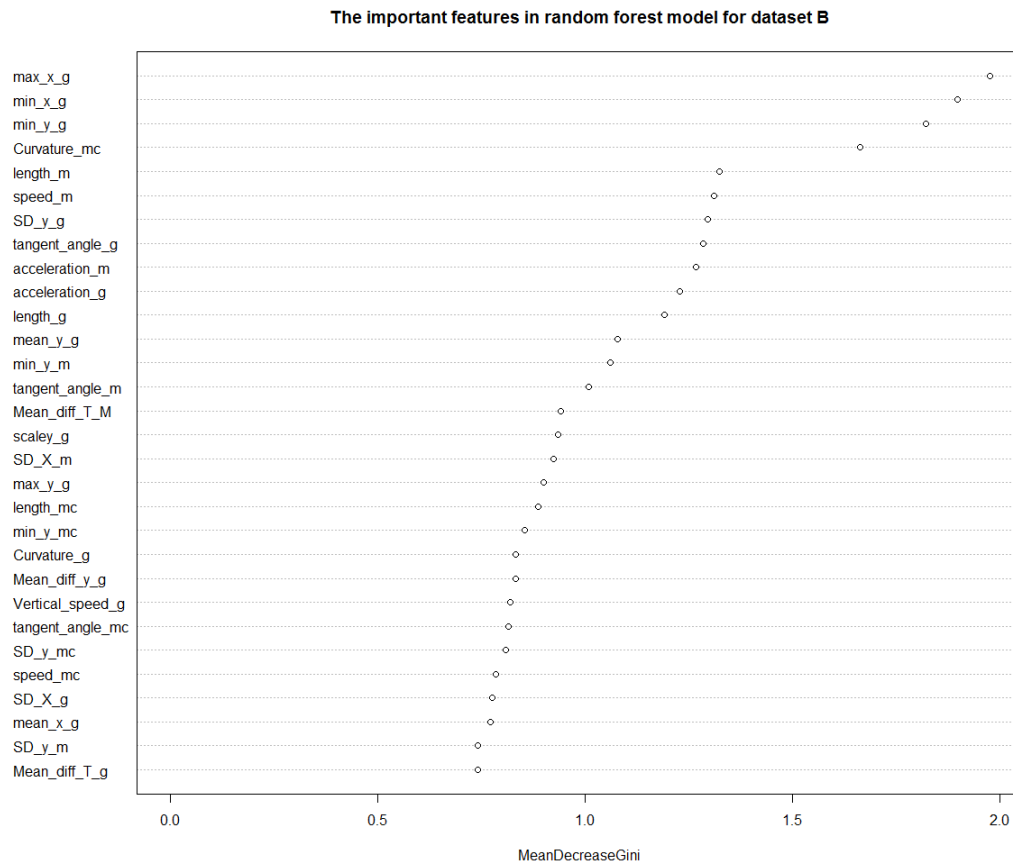
FIGURE 5.2: Important variables in RF model for data B

Figure 5.3 describes the importance ranking for RF variables which built on data set C. The most important variables are the minimum eye movements on y coordinates, the mean of eye movements on y coordinates and the maximum eye movements on x coordinates. Then, speed and acceleration of mouse click movement are also important.
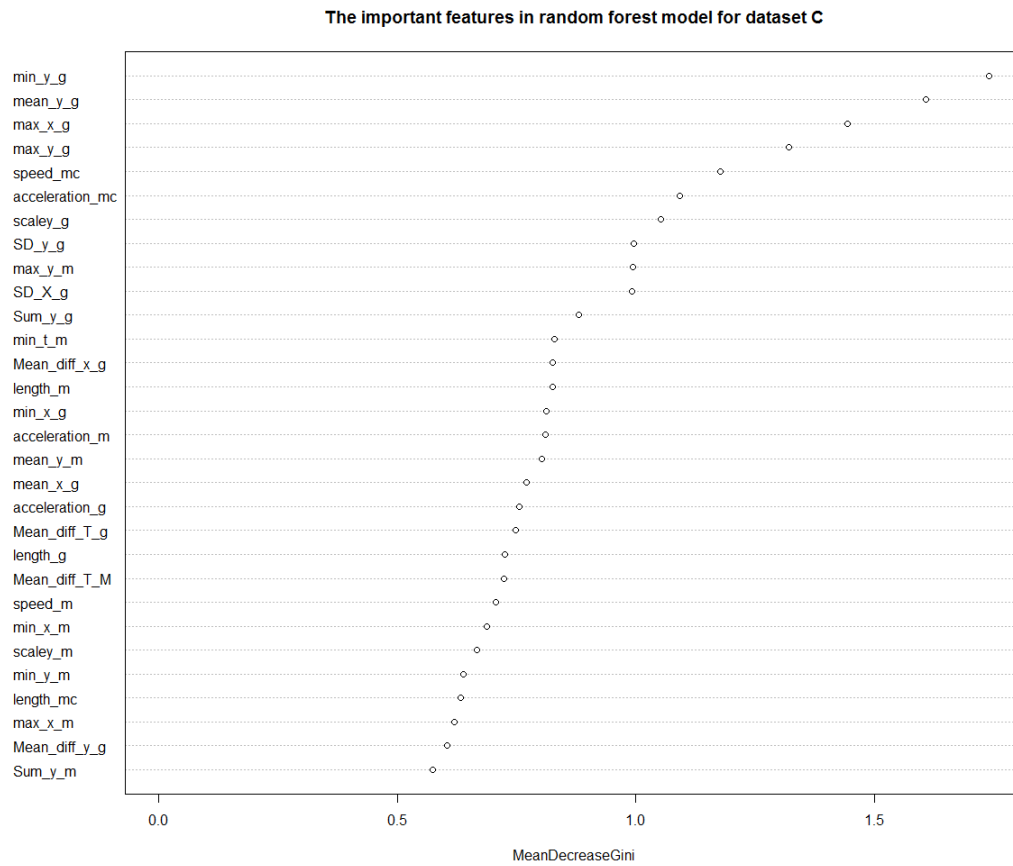
FIGURE 5.3: Important variables in RF model for data C

Table 5.2 shows the results for the three different time sessions of the data sets by considering one session as test data (27 observations) and the other two session as train data (54 observations) then applying random forest model to each data with the RF parameters (mtry=25, ntree=1000). There are slight difference in accuracy rate in each trial for the same data set that due to changing of time which may affect the behavior of the subject through recording the data.

TABLE 5.2: Summary of random forest accuracy for all different train data

Name	First train data accuracy	Second train data accuracy	Third train data accuracy
Data A	0.568	0.568	0.593
Data B	0.389	0.367	0.356
Data C	0.389	0.431	0.444

In order to compare the NN neighbor classifier results which are based on leaving one observation out and find the similarity distance for the rest 80 observations. So

in RF approach it is required leave one observation out of bag every time and build the output model for the 80 observations and get the predictions based on the out of bag observation (test). This done by sampling with changing one observation (Test) each time from the total data.

TABLE 5.3: Summary of random forest and Nearest neighbor accuracy for all data sets

Data name	RF accuracy	NN accuracy
Data A	0.605	0.0170
Data B	0.344	0.0222
Data C	0.681	0.0174

Finally, the performance of the nearest neighbor was not satisfactory for all input features or the other extracted ones . It obtained approximately zero accuracy by taking the average of all accuracy from other features as in Table 5.1, while the RF predictions for test data do better than NN results as in Table 5.3.

Chapter 6

Conclusion

Biometric is a technical approach that consist of recognizing people by extracting and measuring their physical and/or behavioral features. In this thesis, a behavioral biometric discipline is proposed that uses mouse and gaze dynamics as authentication system.

This research concentrate on how to identify the users based on their mouse and eye coordinates using recognition algorithms. Users who are required to be authenticated will have to prove their claimed identities. If another user's mouse movement data does not match the authenticated user then the user will be unauthenticated and rejected by the system.

The main conclusions of this research are summarized in this chapter.

In Chapter4 and Chapter5 the data is described and visualized. These datasets have different features length and different number of class quantities. Selected datasets require transformation into equally spaced observations. After that, normalizing time series data and transforming all values into a common scale and same range. This is required for data preparation since normalization makes the KNN algorithm easier to learn.

This study handled three sets of time series data: First dataset A of 27 users, dataset B of 30 users and another set C of 24 users under controlled circumstances from Look and click website[40].

Thesis focused on similarity measures and extracting meaningful features from multiple data sequences. That will enable good performance in classification and similarity queries of time series.

The features that have been extracted from the original data are statistically analyzed like finding: sum, minimum, maximum, mean, standard deviation, difference, speed, acceleration, range and others operations are computed. This analysis produces 74 features for all mouse and eye movements. The classification methods of time series that have been used are KNN and decision trees techniques. The study involves the evaluation of system performance in terms of verification accuracy and time. Multiple experiments are performed using 1-NN classifier with DTW. In addition a random forest framework is approached for mouse and eye movements classification problem. Building random forest model consists of two procedures. Firstly, feature construction which considered the most challenging and time consuming process, because of infinite possibilities of creating feature sets for all mouse and eye movements. After that, training the modified features data to build the model. Chapter5 showed the results for 1NN method and random forest model. The experimental results were competitive in our proposed biometric identification model. The Maximum 60 % accuracy was achieved using RF, in predicting users identities through mouse and eye tracking. While the accuracy from implementing NN classifier was not satisfactory.

The results are obtained by performing all experiments for all data sets and all class subjects to get the total classification performance. That has been calculated by comparing expected target labels with predicted target class labels. The detailed results can be observed from the confusion matrices A.

6.1 Future Work

The drawbacks of this research are that, the collected datasets and number of subjects are not enough to draw good results. So the challenge is to grow up the datasets into large scale in order mimic the real and to get accurate learning model. In this study three data sets are used and the average of subjects for each data set around 27 subjects which is limited number for getting high accuracy. So the plan to increase number of participants in those experiments also increase number of sessions to seven sessions for each subject instead of three. On the other hand number of features that have been used for building the model are 74 not all of

them are effective, the need to extract more valuable and important features that will improve the accuracy of the learning model.

Appendix A

Tables

TABLE A.1: Mouse movements extracted features

Feature Notation	Feature description	Measures
sub	class label	mean
scalex_m	The mean of mouse normalized X coordinates position for all classes	Z=X-$\mu_{\overline{x}}$
scaley_m	the mean of mouse normalized Y coordinates position for all classes	Z=Y-$\mu_{\overline{x}}$
scalet_m	the mean of mouse normalized time for all classes	Z=T-$\mu_{\overline{x}}$
SD_X_m	the mean of standard deviation for mouse X coordinates position for all classes	standard deviation
SD_y_m	the mean of standard deviation for mouse Y coordinates position for all classes	standard deviation
SD_t_m	the mean of standard deviation for time mouse movements for all classes	standard deviation
Mean_diff_T_M	the mean of the difference Time between two consecutive points for all classes	difference
Mean_diff_x_M	the mean of the difference between two consecutive points of X coordinate for all classes	difference
Mean_diff_y_M	the mean of the difference between two consecutive points of Y coordinate for all classes	difference
max_t_m	The maximum value for time mouse movements for each subject	Max
max_x_m	The maximum value for mouse X coordinates position for each subject	Max
max_y_m	The maximum value for mouse Y coordinates position for each subject	Max
min_t_m	The minimum value for time mouse movements for each subject	min
min_x_m	The minimum value for mouse X coordinates position for each subject	min
min_y_m	The minimum value for mouse Y coordinates movements for each subject	min
mean_t_m	The average time mouse movements for each subject	mean
mean_x_m	The average for mouse X coordinates for each class	mean
mean_y_m	The average for mouse Y coordinates for each class	mean
Sum_t_m	The summation for all time movements for each subject	sum
Sum_x_m	The summation for all X coordinates positions for each subject	sum
Sum_y_m	The summation for all Y coordinates positions for each subject	sum
speed_m	The ratio of the total distance that the mouse traveled from one point to another	
length_m	Length of the curve is the sum of the distances between all adjacent curve co-ordinates for each subject.	length(c)= $\sum_{i=1}^{n} \sqrt{(X_i - X_{i-1})^2 + (Y_i - Y_{i-1})^2}$
acceleration_m	The acceleration of the mouse movements which computed by finding the speed divided by the total time difference between continuous two points.	Δv Δt

TABLE A.2: Mouse click movements extracted features

Feature	Feature description	measures
sub	class label	mean
scalex_mc	The mean of mouse click normalized X coordinates position for all classes	Z=X-$\mu_{\overline{x}}$
scaley_mc	the mean of mouse click normalized Y coordinates position for all classes	Z=Y-$\mu_{\overline{x}}$
scalet_mc	the mean of mouse click normalized time for all classes	Z=T-$\mu_{\overline{x}}$
SD_X_mc	the mean of standard deviation for mouse click X coordinates position for all classes	standard deviation
SD_y_mc	the mean of standard deviation for mouse click Y coordinates position for all classes	standard deviation
SD_t_mc	the mean of standard deviation for time mouse click movements for all classes	standard deviation
Mean_diff_T_Mc	the mean of the difference Time between two consecutive points for all classes	difference
Mean_diff_x_Mc	the mean of the difference between two consecutive points of X coordinate for all classes	difference
Mean_diff_y_Mc	the mean of the difference between two consecutive points of Y coordinate for all classes	difference
max_t_mc	The maximum value for time mouse click movements for each subject	Max
max_x_mc	The maximum value for mouse click X coordinates position for each subject	Max
max_y_mc	The maximum value for mouse click Y coordinates position for each subject	Max
min_t_mc	The minimum value for time mouse click movements for each subject	min
min_x_mc	The minimum value for mouse click X coordinates position for each subject	min
min_y_mc	The minimum value for mouse click Y coordinates movements for each subject	min
mean_t_mc	The average time mouse click movements for each subject	mean
mean_x_mc	The average for mouse click X coordinates for each class	mean
mean_y_mc	The average for mouse click Y coordinates for each class	mean
Sum_t_mc	The summation for all time mouse click movements for each subject	sum
Sum_x_mc	The summation for all X coordinates mouse click positions for each subject	sum
Sum_y_mc	The summation for all Y coordinates mouse click positions for each subject	sum
speed_mc	The ratio of the total distance that the mouse click traveled from one point to another divided by the total time taken to complete the movement.	
length_mc	Length of the curve is the sum of the distances between all adjacent curve co-ordinates for each subject.	length(c)= $\sum_{i=1}^{n} \sqrt{(X_i - X_{i-1})^2 + (Y_i - Y_{i-1})^2}$
acceleration_mc	The acceleration of the mouse click movements which computed by finding the speed divided divided by the total time difference between continuous two points.	Δv Δt

TABLE A.3: Random forest confusion matrix for data A

	s001	s006	s007	s008	s011	s012	s014	s015	s016	s017	s018	s019	s020	s021	s023	s024	s026	s028	s031	s032	s033	s034	s036	s042	s043	s044	s045	class.error
s001	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s006	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.33333333333333298
s007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	0	1
s008	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0.33333333333333298
s011	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1
s012	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1
s014	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.33333333333333298
s015	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.66666666666666696
s016	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1
s017	0	0	1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.33333333333333298
s018	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0.33333333333333298
s019	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s020	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0.33333333333333298
s021	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s023	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
s024	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0.33333333333333298
s026	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0.33333333333333298
s028	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0.66666666666666696
s031	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0
s032	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0.33333333333333298
s033	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0.66666666666666696
s034	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0
s036	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
s042	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0
s043	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0.33333333333333298
s044	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0
s045	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1

TABLE A.4: Random forest confusion matrix for data B

	s001	s002	s003	s007	s010	s011	s012	s013	s014	s015	s016	s018	s020	s021	s022	s023	s024	s025	s027	s030	s031	s033	s035	s037	s038	s039	s040	s041	s042	s043	class.error
s001	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s002	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
s003	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0.33333333333333298
s007	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.33333333333333298
s010	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1
s011	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s012	0	0	0	0	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.33333333333333298
s013	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0.33333333333333298
s014	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
s015	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
s016	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
s018	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0.33333333333333298
s020	0	0	0	0	0	0	0	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.33333333333333298
s021	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0.33333333333333298
s022	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0.33333333333333298
s023	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0.66666666666666696
s024	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.66666666666666696
s025	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0.66666666666666696
s027	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0.66666666666666696
s030	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
s031	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0
s033	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
s035	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0.66666666666666696
s037	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0.66666666666666696
s038	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
s039	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1
s040	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0.66666666666666696
s041	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
s042	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0
s043	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1

TABLE A.5: Random forest confusion matrix for data C

	s001	s002	s003	s004	s005	s006	s007	s008	s009	s010	s011	s012	s013	s014	s015	s016	s017	s018	s019	s020	s021	s022	s023	s024	class.error
s001	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s002	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.33333333333333298
s003	0	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.33333333333333298
s004	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
s005	0	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.33333333333333298
s006	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.66666666666666696
s007	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0.33333333333333298
s008	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
s009	0	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.33333333333333298
s010	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1
s011	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0.66666666666666696
s012	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	0.33333333333333298
s013	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	0.33333333333333298
s014	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0
s015	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1
s016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0.66666666666666696
s017	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0.33333333333333298
s018	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0.33333333333333298
s019	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0.66666666666666696
s020	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1
s021	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0.33333333333333298
s022	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2	1
s023	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1
s024	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	0	1

Bibliography

[1] A.Salaiwarakul. *Verification of secure biometric authentication protocols.* PhD thesis, University of Birmingham, 2010.

[2] L.Ma, C.Yan, P.Zhao, and M.Wang. A kind of mouse behavior authentication method on dynamic soft keyboard. In *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on*, pages 000211–000216. IEEE, 2016.

[3] A.Fustier and V.Burger. Assignment 1 biometric authentication. *Internet Security and Privacy2G1704*, (1), 2005.

[4] K.Adhatrao, A.Gaykar, R.Jha, and V.Honrao. A secure method for signing in using quick response codes with mobile authentication. *arXiv preprint arXiv:1310.4000*, 2013.

[5] P.Kasprowski and J.Ober. Eye movements in biometrics. In *International Workshop on Biometric Authentication*, pages 248–258. Springer, 2004.

[6] D.J.Berndt and J.Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

[7] Eye Tracking Update. Top 7 eye tracking industry developments every investor must know.

[8] A.M.Castillejos. *Management of time series data.* PhD thesis, University of Canberra, 2006.

[9] P.Esling and C.Agon. Time-series data mining. *ACM Computing Surveys (CSUR)*, 45(1):12, 2012.

[10] C.A.Ratanamahatana, J.Lin, D.Gunopulos, E.Keogh, M.Vlachos, and G.Das. Mining time series data. In *Data mining and knowledge discovery handbook*, pages 1049–1077. Springer, 2009.

[11] T.C.Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.

[12] and M.Stamenković M.Milanović. Data mining in time series. *Ekonomski horizonti*, 13(1):5–25, 2011.

[13] C.Kleist. *Time Series Data Mining Methods: A Review.* PhD thesis, Humboldt-Universität zu Berlin, 2015.

[14] B.Sayed. *A static authentication framework based on mouse gesture dynamics.* PhD thesis, University of Victoria, 2009.

[15] I.Batal and M.Hauskrecht. A supervised time series feature extraction technique using dct and dwt. In *Machine Learning and Applications, 2009. ICMLA'09. International Conference on*, pages 735–739. IEEE, 2009.

[16] C.Özgen. *Classification of Forest Areas by K Nearest Neighbor Method: Case Study, Antalya.* PhD thesis, Citeseer, 2008.

[17] S.Lenser and M.Veloso. Non-parametric time series classification. In *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, pages 3918–3923. IEEE, 2005.

[18] J.Grabocka. Invariant features for time-series classification. 2016.

[19] A.Zherdin. *Efficient data mining algorithms for time series and complex medical data.* PhD thesis, lmu, 2016.

[20] K.Fuchs, J.Gertheiss, and G.Tutz. Nearest neighbor ensembles for functional data with interpretable feature selection. *Chemometrics and Intelligent Laboratory Systems*, 146:186–197, 2015.

[21] N.Zheng, A.Paloski, and H.Wang. An efficient user verification system via mouse movements. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 139–150. ACM, 2011.

[22] Miao Liu, Mingjun Wang, Jun Wang, and Duo Li. Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and chinese vinegar. *Sensors and Actuators B: Chemical*, 177:970–980, 2013.

[23] A.Joshi, C.Monnier, M.Betke, and S.Sclaroff. A random forest approach to segmenting and classifying gestures. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–7. IEEE, 2015.

[24] C.Shen, Z.Cai, and X.Guan. Continuous authentication for mouse dynamics: A pattern-growth approach. In *Dependable Systems and Networks (DSN), 2012 42nd Annual IEEE/IFIP International Conference on*, pages 1–12. IEEE, 2012.

[25] F.Betances, A.Pine, G.Thompson, H.Zandikarimi, and V.Monaco. Mouse biometric authentication. *Proceedings of Student-Faculty Research Day, CSIS, Pace University, New York, May 2nd*, 2014.

[26] J.Toenyes, S.Vange, and S.Chaudhry. User gaze predictions in web browsers from mouse movement events.

[27] D.J.Liebling and S.T.Dumais. Gaze and mouse coordination in everyday work. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 1141–1150. ACM, 2014.

[28] and M.Cochinwala D.Bassu and A.Jain. A new mobile biometric based upon usage context. In *Technologies for Homeland Security (HST), 2013 IEEE International Conference on*, pages 441–446. IEEE, 2013.

[29] C.Song, A.Wang, K.Ren, and W.Xu. Eyeveri: A secure and usable approach for smartphone user authentication. In *Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on*, pages 1–9. IEEE, 2016.

[30] Y.Aksari and H.Artuner. Active authentication by mouse movements. In *Computer and Information Sciences, 2009. ISCIS 2009. 24th International Symposium on*, pages 571–574. IEEE, 2009.

[31] C.Shen, Z.Cai, X.Guan, Y.Du, and R.A.Maxion. User authentication through mouse dynamics. *IEEE Transactions on Information Forensics and Security*, 8(1):16–30, 2013.

[32] D.Hema and S.Bhanumathi. Mouse behaviour based multi-factor authentication using neural networks. In *Circuit, Power and Computing Technologies (ICCPCT), 2016 International Conference on*, pages 1–8. IEEE, 2016.

[33] M.Martin, T.Marija, and Arsenovski A.Sime. Eye tracking recognition-based graphical authentication. In *Application of Information and Communication Technologies (AICT), 2013 7th International Conference on*, pages 1–5. IEEE, 2013.

[34] J.Weaver, K.Mock, and B.Hoanca. Gaze-based password authentication through automatic clustering of gaze points. In *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, pages 2749–2754. IEEE, 2011.

[35] S.Easwaramoorthy, F.Sophia, and A.Prathik. Biometric authentication using finger nails. In *Emerging Trends in Engineering, Technology and Science (ICETETS), International Conference on*, pages 1–6. IEEE, 2016.

[36] S.Acharya, A.Fridman, P.Brennan, P.Juola, R.Greenstadt, and M.Kam. User authentication through biometric sensors and decision fusion. In *Information Sciences and Systems (CISS), 2013 47th Annual Conference on*, pages 1–6. IEEE, 2013.

[37] L.Vanamo. Player authentication based on mouse and keyboard dynamics. *Computer Science*, 2016.

[38] Y.Zhang. *Biometric Verification of a Subject Based on Data Mining of Saccade Eye Movement Signals.* Tampere University Press, 2014.

[39] P.Kasprowski and K.Harezlak. Fusion of eye movement and mouse dynamics for reliable behavioral biometrics. *Pattern Analysis and Applications*, pages 1–13.

[40] C.Chen.

[41] P.U.Hattipoğlu. *Time series classification using deep learning.* Middle East Technical University, 2016.

[42] A.Weiss, A.Ramapanicker, P.Shah, S.Noble, and L.Immohr. Mouse movements biometric identification: A feasibility study. *Proc. Student/Faculty Research Day CSIS, Pace University, White Plains, NY*, 2007.

[43] A.Shah, M.Warren, R.Zack, and C.Tappert. Keystroke biometric authentication system experimentation.

[44] T.Kinnunen, F.Sedlak, and R.Bednarik. Towards task-independent person authentication using eye movement signals. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, pages 187–190. ACM, 2010.