Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

Oliver Dagsland Tverrå

Continuous Determination of Age and Gender

Master's thesis in Master in Information Security
Supervisor: Patrick Bours
Co-supervisor: Estelle Cherrier
June 2023

NTNU
Norwegian University of
Science and Technology

Oliver Dagsland Tverrå

Continuous Determination of Age and Gender

Master's thesis in Master in Information Security
Supervisor: Patrick Bours
Co-supervisor: Estelle Cherrier
June 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Dept. of Information Security and Communication Technology

NTNU
Norwegian University of
Science and Technology

Continuous Determination of Age and Gender

Oliver Dagsland Tverrå

June 1, 2023

Acknowledgements

I would like to thank my supervisors Estelle Cherrier from ENSICAEN and Patrick Bours from NTNU. Considering that a large part of the study is for the author to perform research and experiment, they were fantastic in keeping my motivation up and providing guidance towards the study, whilst also making time for meetings and discussions. I would also like to thank those who participated in the data gathering task, enabling the performance of this study as intended. Finally, I would like to thank my family, friends and peers at NTNU for discussing and motivating me throughout.

Abstract

When someone communicates online with another person, there are many aspects that is not known in regards to these parties. The individuals that communicate may also have malicious intentions, and as a result may lie in regards to age and gender. As a result of this, a victim of such intent can suffer consequences like inappropriate pictures being shared online, grooming attempts or physical harm. By utilizing keystroke dynamics, it is possible to determine a human's age or gender. Some studies have achieved varying results in this regard by utilizing keystroke dynamics and communication data. In this study, the determination of age and gender will be the focus; the age and gender groups was divided into 2 classes respectively. However, in a continuous manner, using keystroke dynamics that results in the continuous determination of soft biometrics with keystroke dynamics. The output will be further analyzed, processed and then represented statistically. Research in the area displays promising results when basing the prediction on the full data set and periodic predictions. The earliest determination of gender with satisfactory accuracy achieved for this study was a mean of 126 keystrokes and an accuracy of 71%. The highest accuracy was 87.5% with 1644 keystrokes. For age the earliest was a mean of 312 keystrokes with 72% accuracy, the highest accuracy was 77% with 825 mean keystrokes. There were approximately 1750 keystrokes for each participant, meaning that the earliest determination needed approximately 7% of the participant's writing to determine gender and 17% of the writing in terms of age. Therefore, it is found that it is possible to determine gender and age continuously and early but at the cost of accuracy.

Sammendrag

Når noen kommuniserer digitalt med en annen person, er det mange aspekter som ikke er kjent i forhold til disse partene. Personene som kommuniserer kan også ha ondsinnede hensikter, og som et resultat kan de lyve i forhold til alder og kjønn. Som et resultat av dette kan et offer for en slik hensikt få konsekvenser som upassende bilder som deles på nettet, grooming eller fysisk skade. Ved å bruke keystroke dynamics er det mulig å bestemme et menneskes alder eller kjønn. Noen studier har oppnådd varierende resultater i denne forbindelse ved å bruke keystroke dynamics og kommunikasjonsdata. I denne studien vil fastsettelse av alder og kjønn være i fokus. Imidlertid, på en kontinuerlig måte, ved å bruke keystroke dynamics som resulterer i kontinuerlig bestemmelse av soft biometrics med keystroke dynamics. Resultatet vil bli videre analysert, bearbeidet og deretter representert statistisk. Forskning på området viser lovende resultater når prediksjonen baseres på hele datasettet og periodiske klassifiseringer. Den tidligste bestemmelsen av kjønn med tilfredsstillende nøyaktighet oppnådd for denne studien var et gjennomsnitt på 126 tastetrykk og en nøyaktighet på 71%. Den høyeste var 87,5% nøyaktighet med 1644 tastetrykk. For alder var det tidligste et gjennomsnitt på 312 tastetrykk med 72% nøyaktighet, det høyeste var 77% nøyaktighet med 825 gjennomsnittlige tastetrykk. Det var omtrent 1750 tastetrykk for hver deltaker, noe som betyr at den tidligste klassifiseringen trengte omtrent 7% av deltakerens skriving for kjønn og 17% av deltakerens skriving for alder. Derfor er det funnet at det er mulig å bestemme kjønn og alder kontinuerlig ganske tidlig, men på bekostning av nøyaktighet.

Contents

Figures

Tables

Chapter 1

Introduction

1.1 Topic covered by the project

In attempts to identify users or ensuring access only to certain persons, biometrics has been widely utilized. An example of that can be the smart phones today, it is well known that smart phones can utilize biometric systems that can identify a individual through facial recognition or fingerprint scans. In other words, the focus on physical aspects of the human. However, behavioural biometrics is a domain of biometrics that focus on how humans behave and is not necessarily as intrusive as biometric systems that focus on physical traits mentioned prior [1]. The result of this study would be contribution towards the body of research concerning continuous determination of age and gender through keystroke dynamics, that may in turn provide positive results in terms of determining age and gender of individuals communicating through the keyboard, and can be further utilized based on the intent.

1.2 Keywords

Continuous age determination, continuous gender determination, behavioural biometrics, soft biometrics, keystroke dynamics

1.3 Problem Description

When communicating online, individuals with malicious intent for instance, groomers can change aspects of themselves like their gender or age. This study aims to determine age and gender in a continuous fashion through the typing rhythm behaviour of humans as early as possible.

1.4 Justification, motivation and benefits

It has been proven that when utilizing keystroke dynamics, it is possible to determine characteristics of individuals like for instance age and gender on full texts [2, 3]. However, for this study the aim is to perform such determination early with motivation towards protecting children from harm. With the child being aware of the possible age and gender of the individual the child is communicating with online, could possibly mitigate behaviour that could introduce damage to the child. This study will possibly contribute towards the mitigation of malicious acts towards individuals online. These are well known problems within both the forensics and cybersecurity domains, and efforts towards contribution to research or mitigating these issues are of value.

1.5 Research questions

The research question for this study is specifically:

- **Can a system that can continuously determine age and gender be developed?**

To be able to answer this question, there is a need to perform research to answer the following supporting questions:

- *What methods are there towards creating such a system?*

The domain of keystroke dynamics is vast and contains an extensive number of possible methods that can be applied. Therefore, it is necessary to answer this question to identify methods that can be applied towards continuously determining age and gender.

- *What is a satisfactory decision threshold?*

When determining gender and age as early as possible, there is a need to consider when this determination is satisfactory. In this study's case, it will be after a number of keystrokes.

- *What is satisfactory accuracy?*

Even though this study will attempt to determine gender and age early, it still will need to be performed with satisfactory accuracy otherwise the determination may as well be performed close to chance. The accuracy will possibly conflict with the decision threshold, as the system will have less information the earlier the determination is made.

- *What feature selection methods can be applied?*

Feature choices are important as they will impact a system's performance. However, it is necessary to identify why certain features should be chosen and how this impacts the system.

- *How effective are the chosen methods?*

It is necessary to determine the performance of the methods chosen, as some methods may perform better than others. Furthermore, why this is the case also needs to be identified if there are differences.

- *How effective is the developed system in regards to continuous determination?*

The performance of the system with the applied methods also needs to be determined. Therefore, metrics that will gauge this will be identified and utilised.

The foregoing questions need to be answered to effectively develop such a system while also enabling the ability to gauge the effectiveness of the chosen methods. Including testing the system and the applied methods to allow for the improvement of the system, and as well towards analysing the results.

1.6 Planned contributions

This study will aim to produce a description and a system that can perform continuous soft biometrics determination of age and gender, as well as literature which will address prior research in this regard. Furthermore, to explore the area and possibly improve and implement existing methods towards the aim of the study.

Chapter 2

Background

This chapter aims to give some background to keystroke dynamics, including an introduction to the core topics for this study.

2.1 Keystroke Dynamics

Keystroke Dynamics (KD) revolve around how humans behave through the use of the keyboard and are, in turn, part of behavioral biometrics [2]. More specifically, KD is humans timing information when interacting with the keyboard [4]. This can be, for instance, the timing information between words and letters [4]. An important topic seen throughout studies is the topic of features [2, 5, 6]. The following sections will discuss different types of biometrics.

2.1.1 Behavioural Biometrics

Behavioral biometrics revolves around the way humans behave [7]. For instance, individual x may behave differently than individual y because of a multitude of reasons [8]. Considering this, multiple methods can be used to determine users through their behavior [9–11]. Typical biometrics systems, which can also be referred to as static authentication systems, identify the individual at the start of a session [5]. An example of this can be, for instance, facial recognition or fingerprint recognition [2]. Behavioral biometrics, on the other hand, focuses on how individuals behave either through, for instance, voice or the keyboard [2]. Considering the latter, individual's behavior can be identified through but not limited to keystrokes [12]. Behavior can also be seen with for instance, stylometry which looks at the style of writing [2], or the writing skill [9]. Furthermore, humans behave in different ways, which may result in unique behavior. An example of this unique variable of humans can be related to the timing information when writing on a keyboard or through data collected through mouse usage [10].

2.1.2 Static Keystroke Dynamics

Static keystroke dynamics is used at the very beginning of the point of entry [13]. An example of such a point could be logging into a computer system as a user using a password. Furthermore, it is typically performed when the data that is necessary for classification has been achieved.

2.1.3 Continuous Keystroke Dynamics

There are much that can be done between the start and end of a conversation session in terms of KD [4]. KD in a continuous fashion can be seen as determining aspects of individuals and updating this determination as the subject continues from the start of the session to the end [5]. Static KD is in contrast to this is utilized in the beginning of a session [5]. Periodic KD on the other hand, attempts to determine aspects of a subject in a periodic fashion, this can be by for instance determining after a number of messages [2]. Furthermore, in periodic KD the block of actions can be used to gauge performance whilst for continuous determination it is actions that are performed in for instance a session [14]. Furthermore, when developing a system with classification methods, it is necessary to consider test data and the data the model will be trained upon [14].

2.1.4 Soft Biometrics

Soft biometrics considers the determination of general characteristics; this can be for instance, age, gender, gait or skin color [15]. All living humans have an age and a gender, and the purpose of gathering such information can be many. However, it is not any good on its own, but rather a combination of different features when utilizing these soft biometric features in attempts to identify [15]. The raw feature data is what is gathered before this data is processed into features and is what can typically be gathered from a dataset related to KD [2]. The raw features can consist of data regarding press and release times in that regard [2, 5]. This refers to for instance, when a certain key was pressed and when another key was released. However, by processing these raw features, it is possible to extract features that can be used to gain timing relations of different keys [2]. Gender determination needs to consider the fact that a human's gender may be contributing to different behaviour [2]. For instance, it can be seen that the writing style and timing can vary to such an extent that it is possible to determine the gender with varying accuracy depending on the studies performing them [2]. In terms of age determination, multiple factors can be considered as well. In 1984, the effects on age and skill in typing were considered [16] and are referred to in newer studies attempting to tackle the challenge of predicting age through computational means [9]. Thereafter, an indication of age can be the speed of writing and corrective actions [4]. Considering the foregoing, soft biometrics focuses on general characteristics and does not attempt to identify, for instance, a user in and by itself. To determine such aspects as gender, considerations towards a male can be that

the male class may write faster or slower than a female. Moving further into the topic of soft biometrics, one will find that behavior can be connected to age and gender as there are studies supporting existing differences between the behavior in relation to age and gender [2, 3, 16]. Furthermore, there is a study that has had good results in terms of determining age, while also providing a dataset [3]. They further discuss that their results are preliminary and can be used as a good starting point for further studies [3].

Chapter 3

State of The Art

The state of the art chapter will discuss relevant literature towards the study, and highlight the current best methods related to this. The following sections will also tackle subject matters necessary to properly understand the applicability and purpose.

3.1 Static Keystroke Dynamics

As discussed in Section 2.1, static keystroke dynamics (SKD) analyze the timing information including the typing rhythm of subjects. Concerning this, many research efforts have tackled different aspects [2, 3, 5]. Figure 3.1 shows a typical SKD system. As can be seen, it consists of an enrolment phase where biometric data from the subject are fed into the system; biometric features are then extracted and stored. In the authentication phase, biometric data from the subject is again provided into the system, but this time biometric samples and a biometric application decision.

Something which is seen throughout and is an important subject matter to understand is the performance metrics [2, 10] as these metrics will determine how effective a system is towards its intended purpose.

3.1.1 Data Gathering

This section's intent is to display that datasets are available and provide discussion. Considering that this study will consist largely of analyzing data processed by a system. It is necessary to consider how and what data it is possible to collect and the effect of choosing a certain dataset over the other. It will not necessarily be solely one dataset that needs to be used, as multiple different datasets can be utilized and compared [2]. The data will also need to be experimented upon, using different approaches, while indeed discussing these to identify potential positive or negative sides regarding the foregoing. Therefore, a dataset may come with its limitation and contexts. This can be seen throughout studies that depend on them

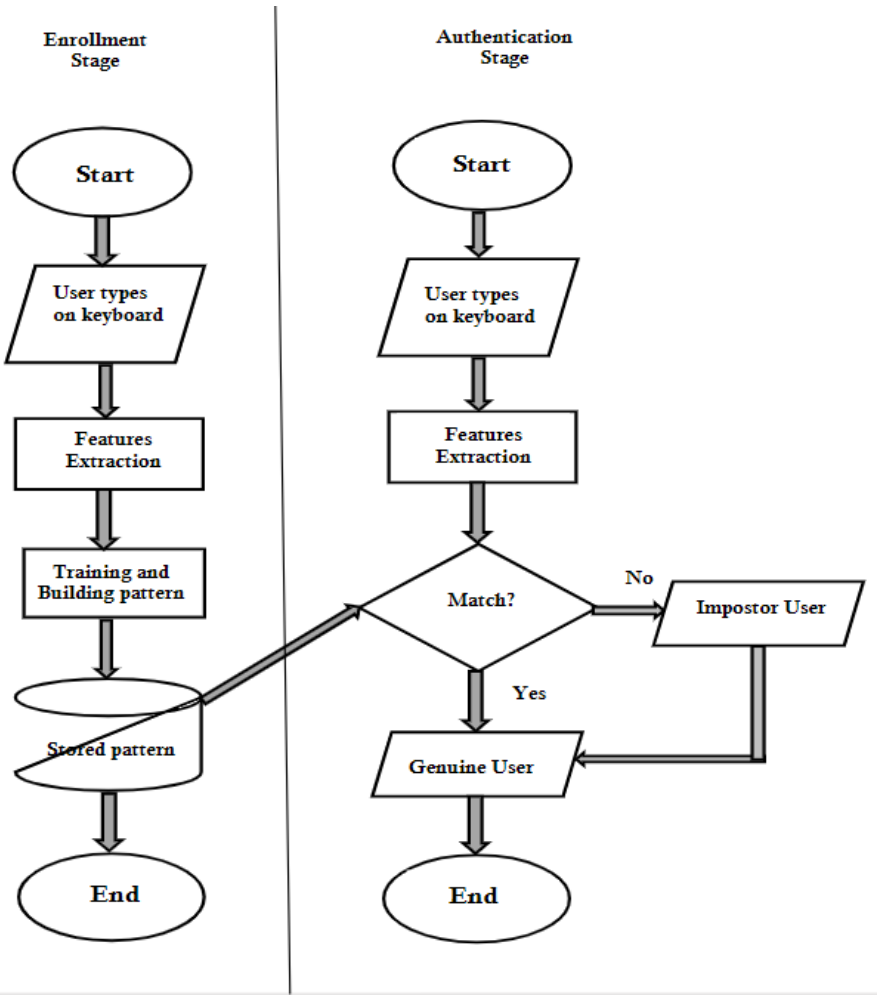


Figure 3.1: SKD System from [13]

[2, 17]. However, there are datasets that are possible to use, as seen in different studies [2–4].

Concerning the content of these datasets, it needs to have the downtime and uptime present as biometric data. This biometric data can then be processed to extract press press latency (PPlat) latency, press release latency (PRlat), release press latency (RPlat), release release latency (RRlat), including the duration of the keystrokes [2]. Biometric data is also labeled with user age, gender, and other available information [3]. To be able to determine the age and gender of a user, it is necessary to have these labels for the subject in the dataset [4]. Otherwise, it would become difficult to ensure that the system works as intended, as without these labels, there is no way of really knowing if the prediction was performed correctly. Language is also necessary to consider, as the user's behavior may differ from one language to another [2]. The keyboard type used is also necessary to consider [4], for instance, if it is a mobile phone, a qwerty keyboard, or another

type of keyboard, as this can also affect the behavior of the user [9] and may vary from a keyboard or for instance a smart phones digital keyboard [18]. Furthermore, as seen in the dataset of [9], they had 997 female participants, 522 male participants making a total of 1519 participants. They also had children from 15 years and below to adults 50 years and above. The dataset was in the Estonian language, and the collection spanned from 2011 until 2017 so they gathered data for a long period. They had multiple data sources; one was where school staff could write in regards to their work, reports, and so forth. Their second source was a questionnaire with 72 questions and 8 free text questions in which they collected data from students that were in the 5th to 12th grade, including parents of the children and staff [9]. They also had other questionnaires in which they collected some more data through. Thereafter, they have a big dataset for their analysis.

3.1.2 Performance Metrics

Starting with the typical ones seen throughout which is False Match Rate (FMR) [5]. FMR is the rate at which comparisons result in a false match. Another metric is False Non Match Rate (FNMR), which is the number of comparisons resulting in a false non match [5]. To identify these error rates, it is necessary to have labels on the data such that the indicators can be counted. Then count the number of times these rates occur. Furthermore, these rates can be used to determine other performance indicators. Typically, static biometric systems performance is measured in FMR and FNMR, resulting in an Equal Error Rate (EER) [5]. Furthermore, the False Rejection Rate (FRR), False Acceptance Rate (FAR), FMR and FNMR are system error metrics that are typically used in biometric systems focusing on biometric recognition [17]. In turn, the performance metrics are derived from comparison attempts towards biometric recognition, and such attempts are performed with the utilization of biometric features and comparison methods [5, 13].

3.1.3 Features

As KD looks at timing information, the typical timing information that it is possible to extract from latency and duration which can be seen in Figure 3.2.

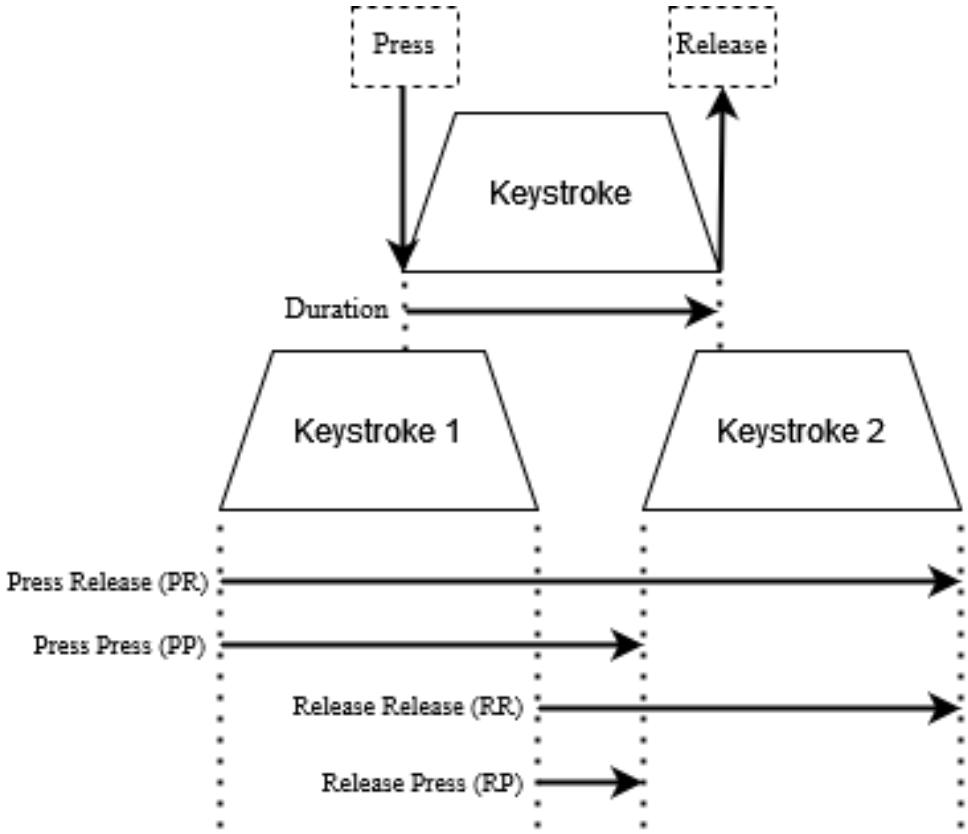


Figure 3.2: KD Timing Features After [17]

The timing information present in Figure 3.2 are common and used in most KD efforts [17, 19]. Furthermore, these are the PRlat, PPlat, RRlat, RPlat, and the duration discussed prior. Latency being the time between keystroke 1 and keystroke 2 for this example. However, the features are not limited to these. It is possible to further define features that can be used in KD like the space duration, backspace duration, arrow keys duration, Shift->I PPlat, Shift->N PPlat [19]. Also, the typing speed can also be used [2, 19]. The foregoing features are simply an example, as there can be a number of combinations of features in KD. Furthermore, efforts using one class classifiers and two class classifiers have been used [13]. The dataset they used consisted of 28 participants, where the participants wrote the prepared text one time each participant. The text consisted of 5000 keystrokes. The predictions performed were to identify users, and the highest accuracy they gained for one class classifiers was by using Histogram Based Outlier Score (HBOS) with an accuracy of 81%, a FAR with 0.20% and FRR of 0.16%. For two class classifiers it was Random Forest (RF) with an accuracy of 91%, FAR: 0.08%, and an FRR of 0.09%. Furthermore, the second place two class classifier was Support Vector Machine (SVM), which gained an accuracy of 90%, FAR of 0.10% and an FRR of 0.09%. This in turn, makes the RF and SVM perform similarly, not concerning the

efficiency. The SVM and RF can be seen throughout literature as the most accurate machine learning methods for SKD [2, 3, 9, 19].

3.2 Keystroke Dynamics - Soft Biometrics

To be able to perform soft biometrics comparisons with KD, it is necessary to identify features that make it possible to determine biometric characteristics of the defined classes [2, 3]. There have been a number of different methods applied, including differing biometric features towards this goal. However, considering results from [1], their findings displayed that it is unnecessary to use a large number of biometric features to achieve accuracy near the system's highest accuracy for determining gender. This, in turn, indicates possibilities to create a system utilizing KD that can determine gender with less training time. The reason is that they used biometric features with high information gain and therefore reduced the number of biometric features necessary to make a decision if the participant was of a certain gender. The lowest amount of biometric features needed when they used the most discriminating features were 50 features, and they received an accuracy for SVM with 73.3% accuracy, 69% accuracy for the naïve Bayes (NB), 77% accuracy for RF. Furthermore, as [19] concludes, is that almost all KD studies the features used were keystroke durations including latencies. Something to note in regards to the foregoing, the number of biometric features necessary before determination does not display the number of keystrokes necessary unless all of the keystrokes are considered. The following section will discuss biometric features and methods that provide the best results regarding soft biometrics.

3.2.1 Soft Biometric Features

The keystroke timing information between keys is what the features consist of, and will then display unique writing characteristics [2]. Timing information used for gender classification can also be latencies and durations. Furthermore, biometric features in terms of age are downtime, latency's, including n-graph latency's and pauses between words [3]. Also, in [9], they indicated that pauses between words could be a good predictor of age. In their research [3] they confirm that keystrokes followed by spaces are of high relevance, as the number one feature they extracted towards multiclass classification was "I -> space", while in third place were "S -> space". Other findings were that there are combinations of keys that have a higher difficulty in terms of typing than other combinations [3]. This is of importance as it displays the need to consider the keyboard that is in use. An example of this can be a Japanese keyboard compared to a qwerty; they are different and can result in different behaviour in terms of writing solely because of the keyboards. The dataset used by [3] was mostly biometric data which has been collected in real life environments, including a large number of participants in the dataset whilst also not excluding biometric data they had available from physical keyboards, phones and tablets even though there were differences with

these modalities. They analyzed 2.3 million keystrokes from approximately 1000 subjects. They also divided the age groups into six different groups. Something also to note is that in [3] they discarded biometric features with a frequency below 500 instances. However, there have been challenges in terms of the dataset throughout literature tackling soft biometrics [2, 3]. Therefore, it is of importance to consider that the dataset will indeed affect the results. Further findings in [3] was that they could not see significant differences between typing on a laptop and desktop computer, considering there were no statements in terms of USB keyboards or the laptop keyboard being used, an assumption towards that a laptop keyboard was what was referred to here. Some features identified for age have been n-graphs latency and pauses between words [3].

As discussed in Section 2.1.4 there exists differences in behaviour in terms of age and gender. The age groups of 20, 40 and 70 years old show differences in typing behaviour including that they are affected to some extent by their mood as well [8]. This was done by looking at the typing accuracy, typing speed, pauses including the variation of typing of the subjects [8]. The foregoing displays that there are indeed differences concerning human behaviour in terms of age and their mood. Thereafter, the problem of determining if an individual is of a certain age becomes more complex as solely a subjects mood may change their behaviour in terms of typing. The duration, latency, n-graph latencies, including the pauses between words has been used in relation to determining age [3]. Another commonality in literature are the algorithms that are used to extract features or to perform prediction, are used for different purposes. Some for early prediction of gender [2], and some for prediction of age [3, 9]. This displays the possibility of implementing methods from studies not from the exact same focus point [5].

Other possibilities in terms of early gender detection with KD is to use KD and stylometry, followed by combining the two with fusion that can be used towards a fused score [2]. The results of their method was a trained biometric model that could make a decision upon the score to determine if the individuals within their dataset were male or female. Their highest accuracy was at 80% by using score fusion with the RF method of classification [2]. An observation [2] made as well were that using only stylometry seemed to give lower accuracy including a lower amount of outliers than when using only KD. Furthermore, the findings in [3] indicated that there are differences in terms of the frequency of the combination of characters that are used in age groups, including that this is correlated to the speed of typing of the respective characters. An example of this is the I followed by space as the number 1 feature for multiclass classification in [3]. Considering the foregoing discussions, there is a high reliance on the dataset including how it is gathered. As a result, the relevance of features can vary in studies. Therefore, the relevance of features will need to be determined and compared. However, to determine general characteristics, it is necessary to apply methods to do so.

3.2.2 Feature Selection

To determine importance of features that are being used, it is necessary to perform feature selection. As discussed prior to this, it can be a high number of features present in KD. The Minimum Redundancy Maximum Relevance (MRMR) algorithm has been used for this purpose, in turn returning a ranking of the features based on the relevance which can be used in the selection of features [2]. In [2], they do this by removing the least useful features with a score equal to 0 as these were the ones containing no useful information. In [1], they encountered approximately over 10000 features and as a result the system got high time complexity. They solved this by calculating the entropy of the system to find the features that reduced the entropy of the system the most with $IG(x, feature) = H(x) - H(x|feature)$ [1]. This method can be seen again in [19]. There are also different methods of separating the probability density function (pdf) of two classes. Furthermore, the Kullback-Leibler divergence (KLD) can be used for that [20]. The KLD is always of a positive value, and the closer this value is to 0 the more similar the pdfs are. This indicates, that the higher the value, the better the feature's predictive power is.

3.2.3 Prediction Methods

When features are selected, and extracted, there is a need to use them towards a comparison decision. One method of doing so, is applying methods such as fused biometric features to gain fused scores from each of the modalities used and can be done by calculating the combined probability scores between 0 and 1. One for male and one for female [2]. Another important utilization of methods was to normalize the data, in [2] they mapped the value of the features within the range of 0 and 1 and did so to remove differences that may affect the prediction. Furthermore, In literature there have been observations where the gender levels determined varied greatly [2]. Therefore, outliers are also important to remove or deal with as they can affect the results.

Furthemore, in [2] they achieved a total accuracy by utilizing fusion with the following domains, a fusion of features accuracy 77% with RF, score fusion by utilizing RF at 80%, KD fusion 78% and stylometry fusion RF 70%. After processing their dataset, they had an average amount of messages per conversation of 54 with an average of 75 keystrokes per message. At the end of the conversation for participants, they had the lowest total accuracy of 60% to the highest of 80% accuracy. They also achieve good accuracy when classifying on 5 messages. However, when an English dataset was used the accuracy fell to the highest with k-NN with 62% by using stylometry fusion [2]. This is understandable, as stylometry and KD are different.

The best results regarding multiclass classification towards determining age was found in [3], and was gained by utilizing RF. They found in their case that RF was a little better than the results gained from C4.5 and SVM. However, the differences were only about 0.02 in f-scores. They used f-score instead of for in-

stance accuracy to gauge performance. Further conclusion in [3] was that building biometric models for a portable device and computers respectively will increase accuracy. The reason for this can be that a keyboard is different than a tablet's virtual keyboard, or a button phone's buttons. Thereafter, the timing information extracted can differ even among the same individual on these different devices. As this utilized machine learning, they needed to balance their dataset before creating the biometric models. The method they used were to undersample the majority class in binary classification and the majority classes in multiclass classifications [3].

It is referred to in [17] towards that proving that algorithms used in KD related to the authentication of users are not biased towards a number of demographics. These demographics are females or young individuals like teenagers and consider this positive towards efforts within behavioral biometrics. In [21], they also discuss the evidence towards that age and gender may affect typing characteristics. This has further been confirmed as well in a number of newer research efforts already [2, 3, 19]. However, the bias that can affect such systems seems to not be discussed to an extent [2, 3]. However, the performances can be compiled that RF displays higher accuracy in studies, however, not by a big margin [2, 3]. RF can be used for multi-class classification or towards continuous responses [22]. Thereafter, it seems to be well suited for continuous prediction. However, SVM has high accuracy but is typically bested by RF in multiple studies [2, 3]. Also, in [2] the K-NN achieved the best performance in terms of accuracy loss in most of their test cases. The SVM can indeed be used for classification [2], and consists of different types of SVM depending on the context of use like for instance Hard-Margin SVM and Multiclass-SVM [23]. The C4.5 machine learning used in [3] for age prediction, has not much worse accuracy than SVM nor RF. However, the C4.5 algorithm can be used to determine if a subject belongs to a certain group and supports features that are of a continuous nature [3, 24]. Logistic regression performed slightly worse than RF [3]. Logistic regression can use variables that change and, thereafter, makes it easier to utilize multiple of them [25]. Therefore, it may be used for continuous prediction as well. Furthermore, in [2] there are mentions of participants that do not conform to the machine learning biometric models and thereafter affect accuracy when utilizing trust levels. They propose a possible solution to this where the trust is reset to its original state after x amount of messages. Furthermore, related to participants that do not necessarily conform to the biometric models. In the study [26], they analyze the Doddington Zoo menagerie which revolves around classifying users into categories of animals, these are sheep, and wolves, lambs and goats describing user behavior. Sheep relates to users that can easily be recognized, goats represent users who are difficult to recognize, lambs contain users who are easy to imitate, and wolves which consist of users who can easily imitate others.

In terms of different methods, this is not an extensive list of all the methods in existence, but rather methods that have been used throughout studies and performed the best for their respective purposes [2, 3]. In Table 3.1, a gathering of

different efforts and their methods will be presented as an overview.

Amount	Subjects	Methods	prediction	Categorization	Year	Paper
2.3 million keystrokes	1000	RF, SVM, C4.5, LR	Age Prediction	6 Age groups	2019	[3]
6545 messages	82	RF: 78% SVM: 77%	Gender prediction	Male, Female	2021	[2]
248 log files	75	SVM, RF, NB, RBFN, MLP	Gender Prediction	Male, Female	2018	[1]
985 acquisitions	51	Scaled Manhattan anomaly detection algorithms	Age, Gender	Male, Female, Teens, Young Adults	2021	[17]
387 log files	84	SVM, SL, NB, BNC, RBFN	Age, Gender	Age and Gender groups	2021	[19]

Table 3.1: KD SB Methods

3.2.4 Testing

To test the biometric models and methods that have been used, in terms of predicting age 10-fold cross validation was used [3]. Literature determining gender has also used 10-fold cross validation [2]. To use 10-fold cross-validation they split their data into 10 evenly sized chunks of data. The testing is then performed 10 times, and an average of the results gained from this in the form of f-score values. Other methods include the Monte Carlo cross validation [10]. Here, the dataset is also split into chunks of data by allocating a percentage as testing data and another percentage as training data. The most extensive method of the testing methods mentioned are the k-fold cross validation, where k can be any number, and where the k is used for training and 1 is used for testing. There are also leave one out tests, which includes removing one participant from the dataset for testing [27].

3.3 Keystroke Dynamics - Continuous Authentication

What separates Continuous Keystroke Dynamics (CKD) from SKD, is that CKD is used directly in a continuous fashion on for instance free text [11]. There are periodic methods, and these methods can involve waiting for a number of messages before having the system react [2]. However, for a truly continuous system it is necessary to have the system react on every keystroke. Therefore, when considering CKD, it is necessary to discuss methods for the system to react when the input is presented to the system. This does not mean that periodic keystroke dynamics (PKD) is not of relevance for a CKD system in its entirety. However, PKD can not be used by itself but rather applied to a CKD system. Furthermore, typing characteristics of single letters of a combination of letters, like for instance, two key combinations, combinations of three keys, or n-graphs, can be used [28]. CKD has had many efforts throughout literature [5, 10, 11, 28, 29]. Further difference from SKD is the performance metrics. These are metrics such as ANGA, including ANIA [28]. These metrics determine the average number of impostor actions (ANIA) and average number of genuine actions (ANGA). Other considerations, is another metric of importance for a CKD system, which is how quickly the system can determine an impostor [5]. An example of this can be seen in [2], where they look at how many messages were necessary to determine that a participant

was male or female. They also calculate the accuracy losses and average number of messages when they attempt to tune the system. Furthermore, in [5] they introduce a trust system for CKD. Here they discuss duration and latency as an important value. Another consideration is that there is a requirement for features that are included, in that the frequency of the features needs to be satisfactory to gain a statistically sound standard deviation including mean [5]. If the feature is rarely used, it will not necessarily give enough information to be of use. Also, considering features used, as can be seen in [30] relevant features like duration, latency, and frequency error are mentioned in terms of CA.

3.3.1 Machine Learning Methods

Machine learning methods that have been utilized through literature in CKD and PKD, include k-means clustering method used in [6], kernel ridge regression [6], RF [2, 3, 6], SVM [2, 3, 6], k-nearest neighbour [6] and neural networks [6]. The same types of machine learning methods are also present in soft biometrics efforts as discussed in Section 3.2.3. However, this does not necessarily mean it is possible to utilize these methods similarly. As seen in literature the RF method typically waits for input as sets of data, for instance as 1 message [2] or as the full data [3]. Therefore, it is necessary to consider further how it is possible to develop a CA system. This is because a CA system will react to every keystroke, it will be periodic of nature if waiting x amount of keystrokes or y amount of time for input only [5]. Something to bear in mind as well is that there are different methods of balancing the data for training. In [3], they undersample the majority class to ensure that the majority class is balanced towards the minority class. In terms of machine learning methods, there are also performance measures that can be used to gauge the performance of such a system. As this study is not dealing with rejections or acceptance of users as discussed in Section 3.1.2, metrics such as accuracy, recall, precision, and f1 score can be applied to gauge the performance of the system further [31]. These are calculated as $Accuracy = \frac{TP+TN}{TP+FN+FP+FN}$, accuracy will gauge how many accurate predictions the system makes, $Precision = \frac{TP}{TP+FP}$, precision refers to how well the system can classify the positive class, $Recall = \frac{TP}{TP+FP}$, recall refers to how well the system can classify the positive class and $Fscore = \frac{2*recall*precision}{recall*precision}$ refers to how balanced the system is in terms of the recall and precision [31]. Furthermore, as seen in [2] they process the biometric sample from the test data with the different machine learning algorithms they use, by computing a score defined as $Score = (w1*ml1+w2*ml2+w3*ml3)/(w1+w2+w3)$. This allows them to manage weak modalities, which can be referred to as for instance parts of a dataset.

3.3.2 Neural Networks

Neural networks receive input to variables called neurons, then these neurons depending on the design, give output based on if the output is correct or not. Fur-

thermore, it can perform backpropagation resulting in a new run applying penalties and rewards [32, 33]. This will in turn improve accuracy. However, neural networks typically need more data to be able to properly train themselves than other machine learning methods. In terms of results, in [32] they gain great results in terms of accuracy up to 99.7%. They apply an ensemble approach utilizing 3 keystrokes at a time. The ensemble is to group the classifiers together towards the goal of prediction, and then voting takes place to provide the final prediction [32]. They also discuss the concept of a keystroke stream consisting of the data being considered from the start til the end. This is defined as $KeystrokeStream = E_1, ..., E_f$. The features utilized in their study are the same as the ones discussed prior to this, just with a different naming convention like for instance Inter Keys Interval (IKI) instead of the timing information representing the timing information between 2 subsequent key presses displayed in Figure 3.2. Other literature also mention that they experience lower EER with a higher amount of keystrokes [33].

3.3.3 Statistical Methods

Statistical methods used ranging from updating trust including prediction in CA can be seen in Table 3.2. In [10], they discuss that instance based algorithms

Method	**Paper**
Euclidean distance	[11, 21, 28]
Manhattan distance	[10, 21, 28]
Cosine similarity metrics	[28]
Scaled Euclidean distance	[6, 28]
Scaled Manhattan distance	[5, 21, 28]
Mahalanobis distance	[11, 21, 28]
Bhattacharyya distance	[28]
Hidden Markov biometric model	[28]
Kolmogorov-Smirnov Test	[10, 28]
Bayesian Classification	[28]
Kernel density estimation	[10]
Energy Distance	[10]

Table 3.2: CA Statistical Methods

are used to be able to utilize as much of the information from the keystrokes as possible. They further utilize the Manhattan and Mahalanobis distance towards computing scores. It is also concluded that previous research in terms of free text used pdf-matching, which in turn needs a significant amount of keystrokes [10]. As seen in Table 3.2 it is a wide range of algorithms available. Many of them achieved great results towards their respective usecase within the differing efforts. In [10] KDE and the Energy Distance performed the best individually. However, depending on the testing size and method used the performance varies. The scaled

Euclidean (SED) and scaled Manhattan distance (SMD) are widely used distance measures in KD and achieve good results [21, 28]. The SMD measures the absolute difference between these two data points, and allows for adjusting factors towards scaling. As can be seen in the research discussed in this paragraph, the features used may have more or less importance [2, 10]. Therefore, the scaled versions of both the Manhattan distance including the eucledian distance are good statistical methods. Furthermore, it is also possible to fuse distance metrics. As seen in [10] they fuse metrics from three algorithms to improve accuracy when attempting to authenticate with fewer keystrokes. A result they gained was by fusing Kolmogorov-Smirnov and Kernel density estimation resulting in lower EER when they were testing samples of both 100 and 200 latency's. The SMD can be seen in [5] towards determining the distance between the template and the input, the reason why this distance was used in their study was because of its good performance in terms of EER. However, in CA systems, a typical topic is the trust system, as the system needs to reach a threshold to be able to reach a decision [2, 5, 6, 28].

3.3.4 Trust System

In [5] they discuss the development of a template. The input is then compared to the templates of the users T_i. There are various manners in which it is possible to create templates that the input can be compared against, but common methods include the calculation of the mean and standard deviations of durations and latencies [5]. The base for these calculations, are common in KD as discussed prior to this. Furthermore, the distance from the template can be determined based on the SMD [5].

In [5] they also discuss reductions or increases in terms of the trust system, by assigning 100 as complete trust and 0 as no trust reducing or increasing these by calculating the values displayed below:

$$d_{dur,p} = \frac{|\mu_{dur,p} - t_{dur,p}|}{\sigma_{dur,p}}$$
$$d_{dur,q} = \frac{|\mu_{dur,q} - t_{dur,q}|}{\sigma_{dur,q}}$$
$$d_{lat,pq} = \frac{|\mu_{lat,pq} - t_{lat,pq}|}{\sigma_{lat,pq}}$$

Then to create a single distance value, as seen in [5] they calculate:

$$D = \frac{(d_{dur,p} + d_{dur,q} + d_{lat,pq})}{3}$$

This then becomes the weighted average of the three distance values presented above [5]. This can be used to update the trust systems level. The method is also similar to the fusing of distance features. However, weight is often used in efforts concerning continuous authentication [2]. The D value can be used to update the trust level either up or down and will reward or provide a penalty depending on the comparison to the distance of D towards threshold T [5]. Furthermore, in [2],

the prediction is performed by applying weights to the comparisons in a static or variable manner. This in turn will move the value of the prediction variable towards thresholds faster or slower depending on the configurations. The thresholds that are used can then be used as a decision point [2, 5]. Also, there are outliers that can affect the system performance, and the mean and standard deviations of different biometric features are a possibility for further improvement upon the accuracy of the system [6]. This has also been seen used in [2], as they removed their outliers by calculating the mean and standard deviation followed by removing values above 3 times the standard deviations from the mean. Another method that was used in [2], was that they performed gender classification on the full dataset before attempting to classify the gender early to be able to compare performance. As the classifications on the full dataset were expected to have higher accuracy than the classifications based on messages.

Further update mechanisms are static, variable, and hybrid [2]. Static is a fixed value, variable is an update mechanism that varies, whilst the hybrid mechanism is a combination of the two. The calculations for the non static update mechanisms: $v = \frac{c\sqrt{l}}{100}$. To elaborate on this calculation, v = the score, c = the prediction score, and l is the length of the message. In [2] they use this to grant the score given a weight based on the length of the message, do the square root of the length of the message, and divide the results by 100 to ensure that the score does not give too much of an update to the trust level. Trust systems are common in CA and are part and partial in how the system is designed [28]. An example of such a system can be seen in Figure 3.3.

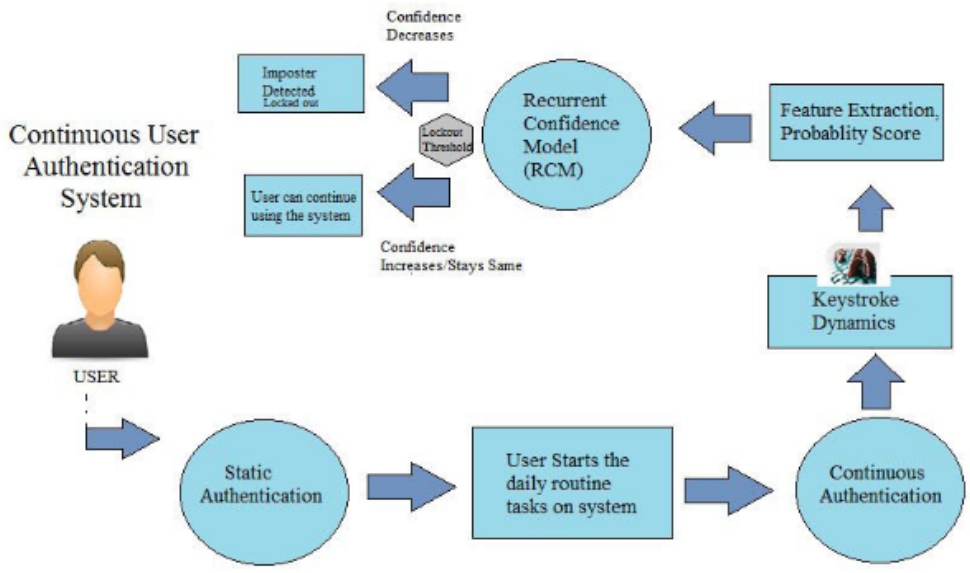


Figure 3.3: CA System from [28]

However, the CA system described in Figure 3.3 stops after the system has reached "The user can continue using the system". The checks should continue

after that point as well. However, as can be seen in Figure 3.3, the data is fed into the system, and the extraction including the biometric model that has been developed is applied resulting in an increase or decrease in the confidence that a user is an impostor or the real user. Also, in Figure 3.3 probability Score can be seen and can typically be derived from machine learning methods [2, 6]. However, in [6] they use dissimilarity scores while similarity scores are discussed as well. The dissimilarity scores are part of their trust biometric model for their CA system and their results are good. Furthermore, there is the possibility to differentiate between unique biometric characteristics between the different classes. For example, if there are in particular biometric characteristics in age groups or gender it is possible to use them as an indicator towards that classification and add more weight [2]. Furthermore, as seen in Figure 3.4, there are methods that allow for combining output from classification methods.

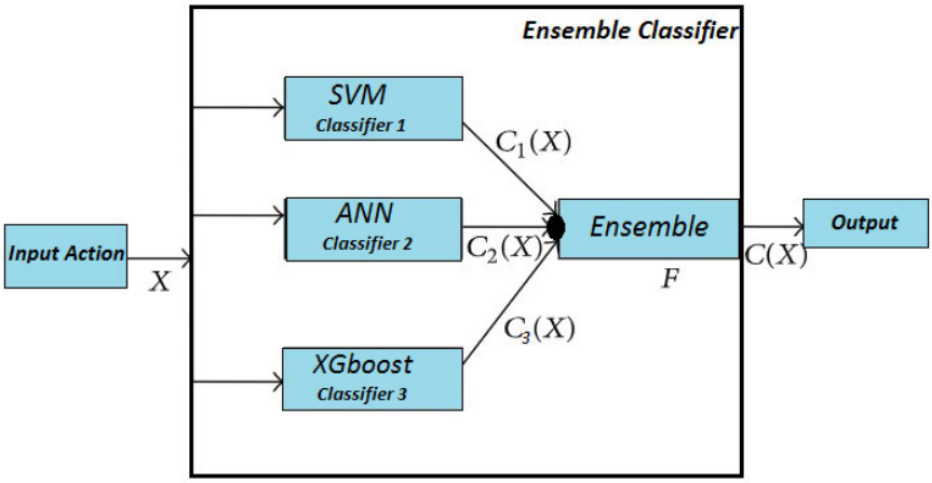


Figure 3.4: Ensemble Method from [28]

A similar approach can also be seen in [2], where they use the RF prediction to create a weight. In general, what has been seen in the literature, is that a CA system needs to allow the users to behave freely to be able to train biometric models or create reference templates such that it works in a "free" environment.

Chapter 4

Data Procurement

This study was comprised of three bigger parts, the data procurement part focused on capturing data which will be used in the study. Then, the data analysis part, and finally, the continuous development of the system, which was ongoing throughout the study. This chapter will discuss the data procurement task.

4.1 Data Capture Task

The data capture task was performed over a time span of 3-4 weeks. This task aimed to perform the biometric acquisition process to gather the main data set. The biometric capture device was the author's laptop, including one specific keyboard that was used for all the data capture tasks in an attempt to ensure that all the participants had an equitable performance for the task. The keyboard was a qwerty Norwegian keyboard, and the participants were asked to write in English. The author did not have the entirety of these weeks available for the data capture task. Therefore, the data capture task activities were split up over time to gain satisfactory data. The motivation behind this specific type of data capture task is that residents in Norway can communicate with other nationalities through online communication platforms or games. Furthermore, the data capture task given to participants was that they were given approximately 10 minutes to write free text. The participants could write about anything they wanted except for sensitive, personally identifiable, or other personal information. Examples given were the universe, a tv-show or a video game. Guidance towards the data capture task was also present in the data capture task software used created by Antoine Jourdan[1]. It can be seen with example text in Figure 4.1.

The program was customized to fit this task, and also to include guidance and reminder of the task for the participants.

[1]Thanks to Antoine Jourdan from ENSICAEN for creating the initial version of the software.

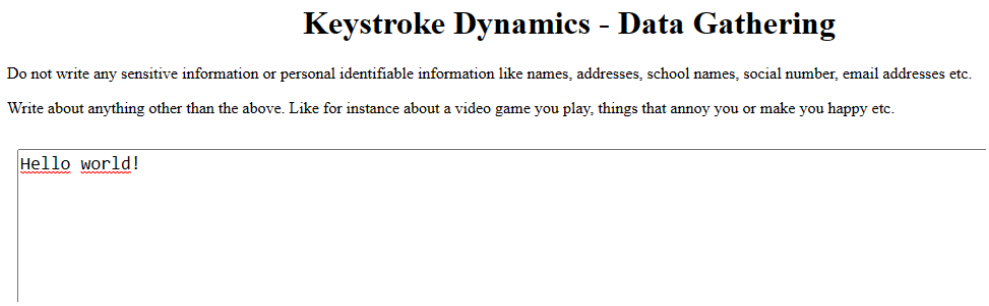


Figure 4.1: Data Capture Task System

4.2 Dataset

The dataset included the following information, which can be seen in Table 4.1.

Column	Meaning
KEYCODE	Unicode representative of a letter typed.
TIMEDOWN	Timing information for key down press.
TIMEUP	Timing information for key release.
LETTER	The string representative of the keystroke.

Table 4.1: Dataset Information

This information was present for each keystroke. An example provided by the author using the data capture task software can be seen in Figure 4.2.

Other information was the age and gender which was necessary towards the purpose of determining the biometric property age and gender of the biometric data subjects continuously. The result of the data capture task is a data set consisting of 56 participants, 44 were less than 30 years of age, while the rest were above 30 years of age. This resulted in approximately 1750 keystrokes each participant, although the real amount of keystrokes the participants wrote varies. In total, in terms of gender, there were 25 female participants and 31 male participants. In Figure 4.3, the spread between the number of keystrokes and the age of the participants can be seen.

KEYCODE	TIMEDOWN	TIMEUP	LETTER
16	1,68387E+12	1,68387E+12	Shift
72	1,68387E+12	1,68387E+12	H
82	1,68387E+12	1,68387E+12	r
69	1,68387E+12	1,68387E+12	e
76	1,68387E+12	1,68387E+12	l
76	1,68387E+12	1,68387E+12	l
8	1,68387E+12	1,68387E+12	Backspace
8	1,68387E+12	1,68387E+12	Backspace
8	1,68387E+12	1,68387E+12	Backspace
8	1,68387E+12	1,68387E+12	Backspace
69	1,68387E+12	1,68387E+12	e
76	1,68387E+12	1,68387E+12	l
76	1,68387E+12	1,68387E+12	l
79	1,68387E+12	1,68387E+12	o

Figure 4.2: Data Example

The spread is due to as mentioned prior, the higher amount of participants in the 20 til 30-year range. The author attempted to balance this when gathering the data. However, this turned out to be quite difficult. This was quite more balanced regarding the number of appearances of male and female, as seen in Figure 4.4. In Figure 4.4, 0 is for females, and 1 is for males. They are not entirely balanced regarding keystroke instances, but that is to be expected as balancing these equally is quite difficult.

The following classes will be created to divide the dataset into groups towards classification. The gender class will be divided groups, the class label 0 is for females and class label 1 for males. The age classes will be determined as follows. The age groups below 27 with class label 0 as (younger adults) and greater or equal to 27 with the class label 1 as (older adults).

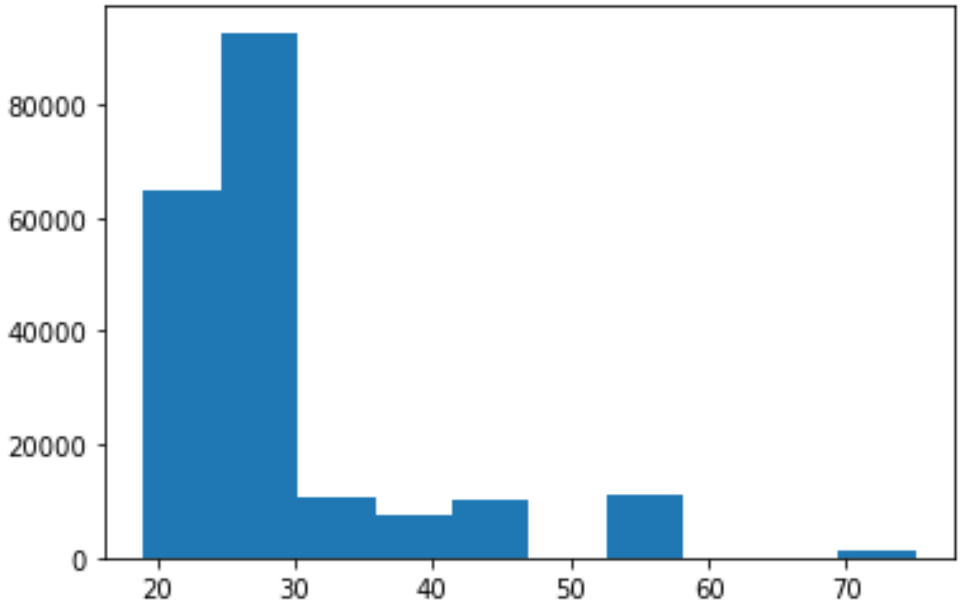


Figure 4.3: Age Spread

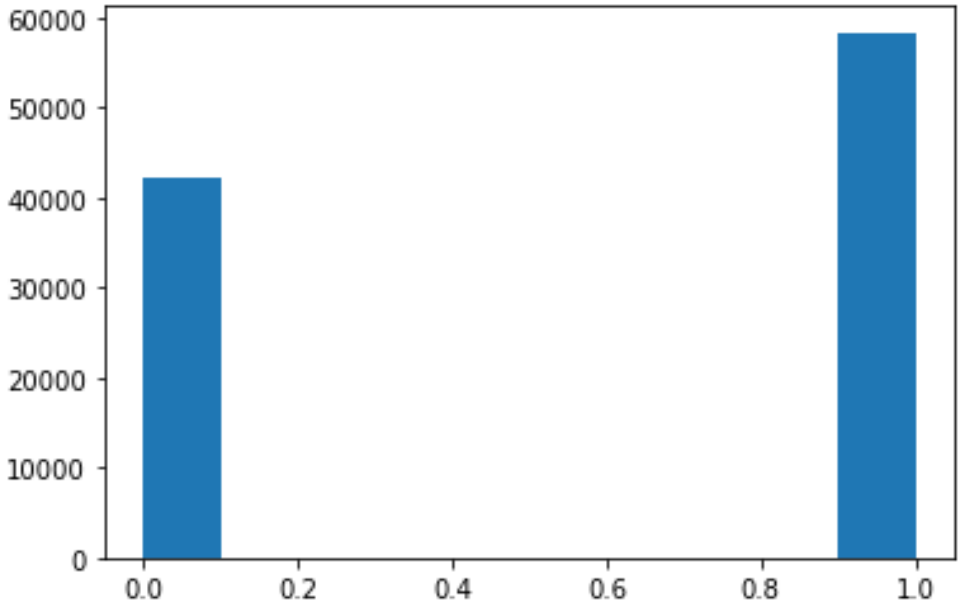


Figure 4.4: Gender Spread

4.3 Dataset Limitations

The data set is limited because it will not represent a high population of residents in Norway. Furthermore, English writing skills can vary greatly, and many studies

in Norway have Norwegian as their primary language of writing in academia. Therefore, writing skills will indeed have an impact as part of the data capture task was performed mainly in an international academic institution. However, there were no purely English or American participants, meaning none had English as their main language. It is also limited because the data was gathered from a specific keyboard with which the participants were unfamiliar. In other words, using a device they are used to will change their writing behavior.

Chapter 5

Analysis

This section will focus on discussing the analysis of the dataset, as well as the resulting system that will be applied in the result chapter. The analysis itself focuses on identifying features which can be used towards determining the classes within the system that is developed. The dataset itself is also discussed as exploring approaches towards identifying the characteristics is necessary.

5.1 Problematic Values

It is necessary to clean the data, specifically there are outliers, errors, or possibly irrelevant data. However, when processing the data it was observed that the data could be represented in a manner that was not understood by the author, the representation of such keys could be the key "Meta" or "Dead" keys which were present in the dataset. However, data could also be lost as some keys were represented as a long string when processing the data. These strings were observed with quite a low number of instances, so much so that they were deemed arbitrary and removed.

5.2 Biometric Features

An important part of the processing of the data, is the processing of the captured biometric sample into biometric features. The extracted biometric features were latency and the durations of the keystrokes. This was performed by utilizing the algorithm displayed below. T1 refers to the timedown, while T2 refers to time up for the first keystroke. T3 refers to timedown and T4 refers to timeup for the second keystroke. Lat_RP refers to the latency for RPlat, Lat_PP for the latency of RRlat, Lat_PR for PRlat latency and finally Lat_PP for PPlat. The DurA is the duration for the first keystroke and DurB is the duration for the second keystroke after DurA.

```
for T1,T2,T3,T4 in raw_data:
    DurA = T2 - T1
```

```
DurB = T4 - T3
Lat_RP = T3 - T2
Lat_RR = Lat_RP + DurB
Lat_PR = DurA + Lat_RP + DurB
Lat_PP = DurA + Lat_RP
```

An example of these processed biometric features can be seen in Figure 5.1. The age and gender, including the participant number, is arbitrary to present and therefore redacted.

Index	Participant	Age	Gender	DurationA	DurationA Key	DurationB	DurationB Key	PR	PP	RR	RP
47549	[redacted]	[redacted]	[redacted]	81	g	79	.	753	674	672	593
47550	[redacted]	[redacted]	[redacted]	79	.	74		418	344	339	265
47551	[redacted]	[redacted]	[redacted]	74		176	Shift	303	127	229	53
47552	[redacted]	[redacted]	[redacted]	176	Shift	72	T	136	64	-40	-112
47553	[redacted]	[redacted]	[redacted]	72	T	70	h	278	208	206	136
47554	[redacted]	[redacted]	[redacted]	70	h	54	e	179	125	109	55
47555	[redacted]	[redacted]	[redacted]	54	e	56	y	139	83	85	29

Figure 5.1: Biometric Features Example

To be able to identify biometric features that are of importance, it is necessary to process the biometric features further. This is performed by doing biometric feature selection and extraction. However, before extracting the biometric features, it is necessary to remove outliers. This process will be discussed in the following section.

5.3 Outlier Removal

Outlier removal was necessary, and it is necessary to consider a method that will not skew the data towards a higher or lower value. Outlier removal was performed by removing values that were 3 standard deviations away from the mean for each biometric feature for the respective classes. An example as to why, is that an outlier in RPlat may indicate that a participant is thinking about what to write. As the data was gathered through a free text task, there were more outliers. It is also well known that humans have different aptitudes in writing, including creativity and how well a human can write freely. The same can be said in terms of durations, if a participant is holding a button for an extended period of time or by accident, the action can result in outliers that will affect the biometric feature value. This can be holding shift for a prolonged period because the participant is unsure where the special characters are placed on the keyboard. Thereafter, affecting the resulting system if not dealt with. Individuals accidentally hit Norwegian keys like "å" or "ø" while also searching for and testing different combinations to locate the correct special character. This is also a result of the participants being able to write free text with an unfamiliar keyboard. For instance, if a participant wants to ask a

question in the text, they can. However, they need to find the combination of keys that makes the question mark, which can result in sporadic use of shifts and special characters. Another limitation of this result in is that this dataset does not reflect participants in their optimal environment. The participant's optimal environment would be the participant's own keyboard.

5.4 Normalization and Balancing

The values within the data set can vary from negative values to reaching positive values as high as 2000 milliseconds. The data was normalized to stay within the range of 0 to 1. Also, considering that the dataset, to some extent, is unbalanced in terms of the age and gender classes. The data which is used to train will undergo randomized undersampling for the machine learning part to ensure that one class is not over-represented. The reason being, is that the participants did not have a specific writing task. Therefore, the single keys or key combinations are not equally represented for the classes.

5.5 Biometric feature Selection and Extraction

biometric feature selection is an important process towards determining which biometric features to include in the biometric template or machine learning biometric models. To be able to determine gender or age, it is necessary to consider biometric features that provide information about the respective classes. As discussed prior, there are an extensive amount of biometric features. Including all of them, can result in a complex system which in turn does not necessarily need to be positive. Furthermore, the information gained from each biometric feature is also necessary to consider. This is important in the systems used, and will therefore not be taken lightly. Considering durations are of important, as they can be used as a biometric feature for every keystroke that is made. In Figure 5.2, the different colors represent the different participants and the black line represents the mean of the total duration. As can be seen, the durations are quite dense in the 0.0 towards 0.2 areas.

However, by separating the data points by gender the following durations can be seen in Figure 5.3. The data points with the reduced size give a clearer picture in the sense that there are indeed some differences although they are quite densely located around the mean of duration.

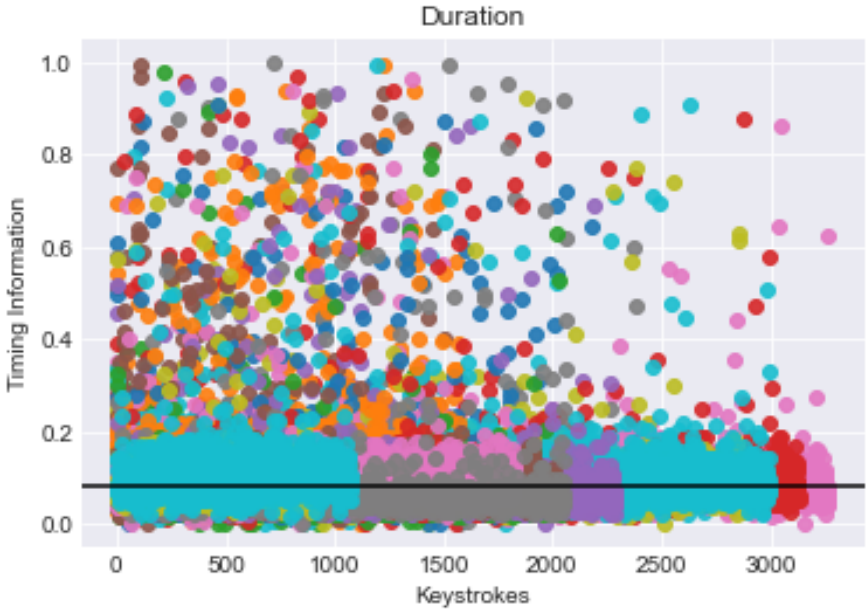


Figure 5.2: Dataset Durations

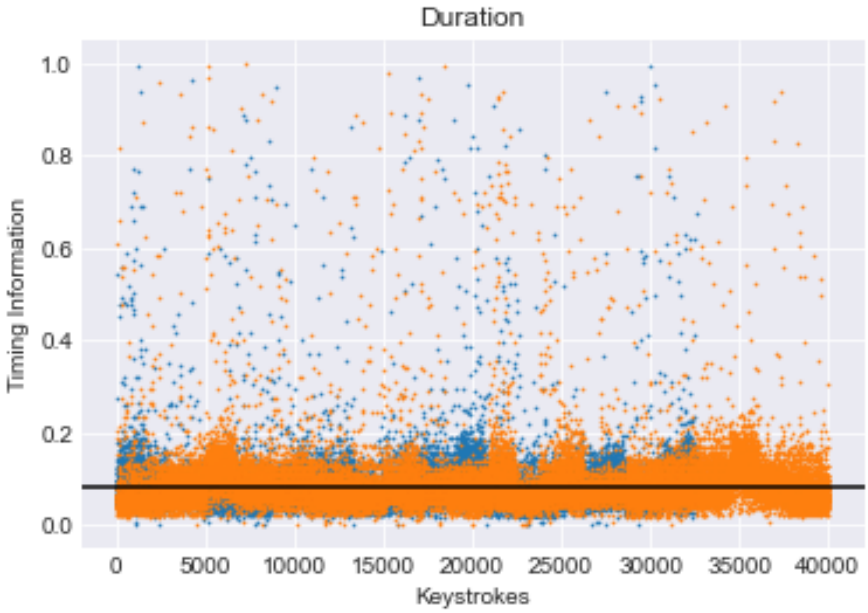


Figure 5.3: Durations Gender

The RPlat is quite more spread out as seen in Figure 5.4. However, further analysis of the RPlat's are needed to better understand them. The data points are also quite densely focused around the mean as in Figure 5.3 but with more spread.

This is because the time from when individuals release and press a key varies, as well as the skill in typing. The reason why the RPlat floats higher above the value 0 on the y axis, is because there can be negative values, as participants can for instance hold shift and then press another key to write a capital letter. It can be seen that this is not necessarily done extensively over longer periods of time, but it seems that multiple participants have done this which can be an indication of an unfamiliar keyboard.

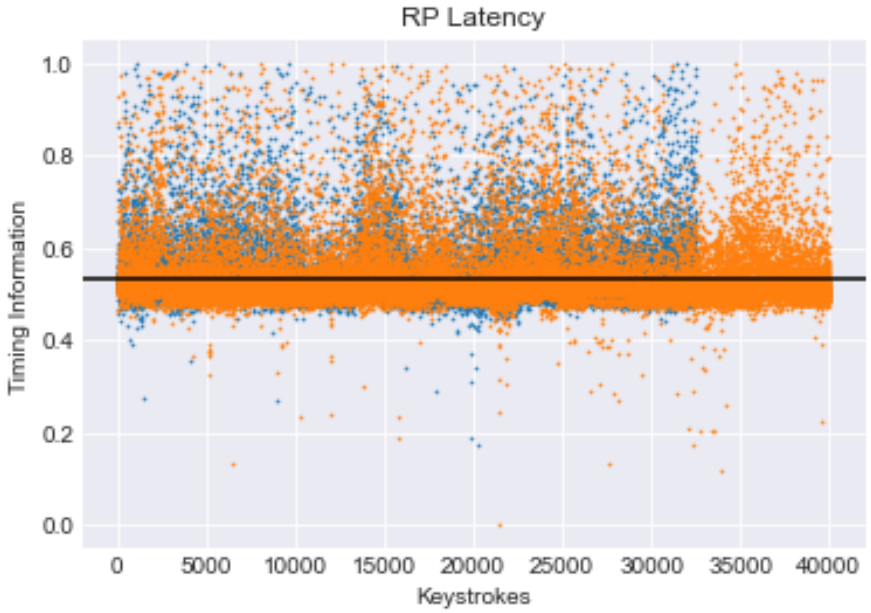


Figure 5.4: RPlat Gender

The PRlat seems to have a higher amount of differences and to be, but this is not necessarily the case. As can be seen, there are some differences on the lower points below the mean, which indicates possible differences.

The final overview of the biometric features which is the PPlat category can be seen in Figure 5.6. As can be seen in contrast to the figures discussed prior, there are some differences in the values below the mean.

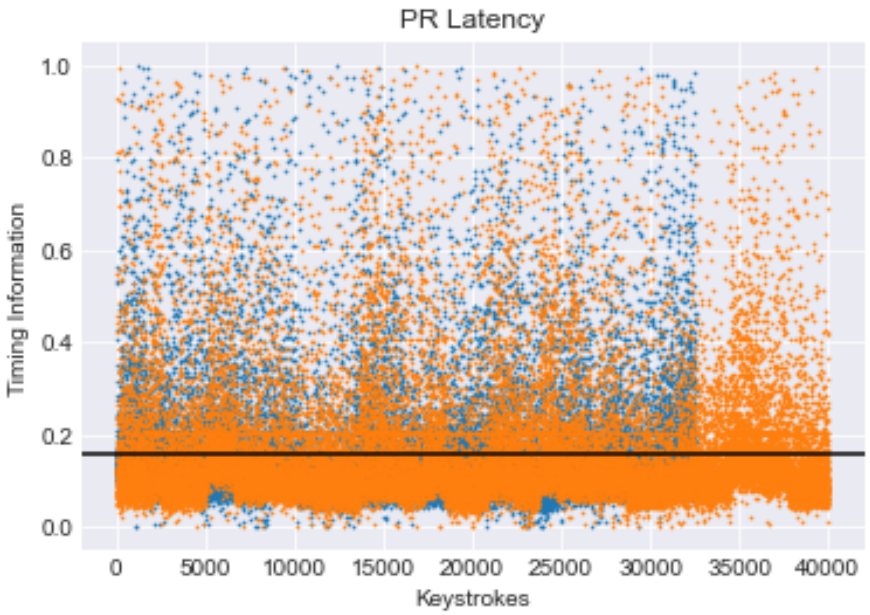


Figure 5.5: PRlat Gender

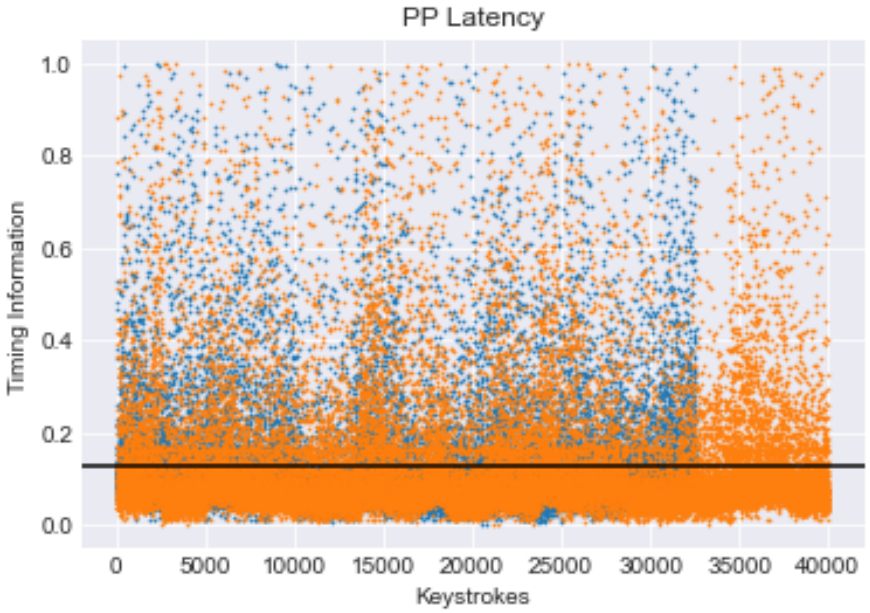


Figure 5.6: PPlat Gender

This overview provided a high-level understanding of the differences in terms of typing. Therefore, it is necessary to analyze individual biometric features further to understand the differences. Further analysis will consist of duration's and

RPlat's, the reason being that the durations are more frequent whilst the RPlat can go in both a positive and negative direction giving a bigger space for the timing information.

5.5.1 Biometric Feature Selection

The selection of biometric features can be performed by generating the pdf of bell curves are is performed to analyse the data. Before applying this, it is necessary to determine the biometric features that are most frequently used. This is important, as the system will react to every keystroke, and attempt to determine as early as possible, it is easier to consider them when an overview of them is presented. These can be seen in Table 5.1 and Table 5.2 for gender. The RS Latency column represents the right side of the keyboard, the LS Latency column represents the left side of the keyboard, whilst the Latencies represent the most used two key combinations. The Duration column represents single keystrokes.

Latencies	Amount	Duration	Amount	RS Latency	Amount	LS Latency	Amount
Backspace Backspace	2838	Space	6946	i n	796	e r	593
e Space	1629	Backspace	4281	o n	358	r e	532
Space t	1328	e	3574	o u	349	a t	392
t h	1057	t	2743	h i	269	s t	353
s Space	1028	a	2322	l l	203	e s	341
Space a	1022	o	2253	l i	196	t e	307
t Space	886	i	2046	o m	157	a r	306
h e	871	s	1829	n o	145	v e	302
i n	796	n	1826	i l	138	a s	276
Space i	680	r	1710	l o	136	s e	239

Table 5.1: Top 10 used key combinations - Female

As can be seen in Table 5.1, the backspace -> backspace latency has most instances by quite the margin. However, they are similar for both gender classes in terms of their distinctiveness, which is quite low. The latency e -> Space RPlat is quite distinct for the gender class. This is a latency where the participant can write the keystrokes with their left hand, for instance, with their index finger and thumb or a combination of their right hand and left hand. Therefore, it does not necessarily mean that there is discriminating value for the biometric features even though they have high instances. As such, analysis of durations and latencies with a number of instances i where $i > 2$ will be performed.

latency	Amount	Duration	Amount	RS latency	Amount	LS latency	Amount
Backspace Backspace	2658	Space	9196	i n	796	e r	593
e Space	1629	Backspace	4989	o n	358	r e	532
Space t	1328	e	4727	o u	349	a t	392
t h	1057	t	3827	h i	269	s t	353
s Space	1028	a	3284	l l	203	e s	341
Space a	1022	o	2979	l i	196	t e	307
t Space	886	i	2916	o m	157	a r	306
h e	871	n	2662	n o	145	v e	302
i n	796	s	2626	i l	138	a s	276
Space i	680	r	2161	l o	136	s e	239

Table 5.2: Top 10 used key combinations - Male

For instance, in keystroke dynamics, one typically does not use corrective actions like "Backspace Backspace" as seen in Table 5.2. An observation of the data, is that generally the female class seem to write slower than the male class. The difference is, however, not to the extreme. Other than the duration's themselves the right side, and left side of the keyboard, including the combinations of keystrokes that were typically used when writing in English, like “in” and “th”, “he” were analyzed to see if these combinations were distinct. The latter combinations are the key combinations for writing the word “the”, which in turn makes these combinations frequently used, which can also be seen in Table 5.2 as they are part of the most used key combinations. However, when looking at the “in” RPlat, they were quite similar between the gender classes. This can be because both classes utilize the same hand in a similar manner when they are typing this combination. This may also be connected to age; in terms of the age classes, most participants were below 30 years of age. However, there were differences in terms of the key combination of “li”. Individuals may write this key combination with their index finger, while some may have different approaches to writing this combination of keys, for instance, by the use of one hand or both hands. The key combinations where participants had to rearrange their hands to type the keys displayed the most distinct biometric features in general. However, there was not much distinctiveness for the “th” RPlat in terms of the age groups, whilst this biometric feature is more distinct for the gender classes, which indicates that the biometric features may differ depending on the gender and age groups. Therefore, it is necessary to consider which biometric features should be extracted to determine age and gender.

The original analysis of the data included analyzing bell curves based on the pdf of the biometric features for the respective classes. This method took quite long to perform. Therefore, the biometric feature selection method that was applied later in the study was the following. Firstly, the KLD is calculated from the biometric feature pdf for the two classes. This is done by utilizing the Python library scipy.stats [1] and is performed as follows:

```
KLD = entropy(pdf_class1, pdf_class2)
```

[1] https://docs.scipy.org/doc/scipy/reference/stats.html

The pdf_class1 stands for the pdf for class 1, for instance, male whilst class 2 is then for female. The probability of the biometric features in the dataset is also calculated. This is done by FI/TDI. In this case, the FI stands for the instances of the biometric feature, and the TDI is the total amount of the biometric feature type, for instance, durations. An example of this will be given below in Table 5.3 to clarify. The higher the value of KLD the better the biometric feature separates the two respective classes; in other words the more distinct the biometric feature is. Furthermore, if the value received back is infinity, then the biometric features pdf does not overlap and, therefore, is highly distinct, providing high predictive power. The lower the value is, the poorer the biometric feature is to separate the classes. The probability refers to the probability that one will encounter the biometric feature in the dataset, in other words, how much the participants use the biometric feature. In this example, the best biometric feature in terms of accuracy will be the latency "Hi" in Table 5.3. However, for more frequent updates, the latency "ur" is good, but if used will come at the cost of accuracy.

Biometric Feature	KLD	Probability
ds	0.0054	0.0008
ur	0.0093	0.0027
Hi	1.4998	0.0001

Table 5.3: Latency Biometric Feature Selection - Gender

Another example of this biometric feature selection method can be seen in Table 5.4 which focuses on selecting biometric features in terms of the duration. The same process as in Table 5.3 applies. Here, the biometric feature "s" is highly likely to be encountered in the dataset. Also, as seen in the KLD, it is approaching the value 0, which indicates that the biometric feature "s" for the male and female class is quite similar to each other. Therefore, the accuracy will be reduced if this biometric feature is included in the reference template. However, the system will have more frequent updates as the probability is higher. The best biometric feature for accuracy, in this case, is the biometric feature "x". The same process is performed for both the gender and age classes.

Biometric Feature	KLD	Probability
s	0.0308	0.0449
x	0.1286	0.0013
P	0.0563	0.0003

Table 5.4: Duration Biometric Feature Selection - Gender

5.6 Testing

The main testing method will be the leave one out test. This will be performed to ensure that the whole dataset is utilized. This will result in 56 tests for each run, with the system attempting to determine the gender including age continuously for every participant.

5.7 System Architecture

To be able to test different approaches, a system was to be made towards this purpose in Python [2]. An overview of the system can be seen in Figure 5.7. Training data refers to the data used for training. Depending on if 1 test or n tests are being performed, the process displayed in Figure 5.7 can be performed n times. The system allows for creating of both biometric templates and machine learning

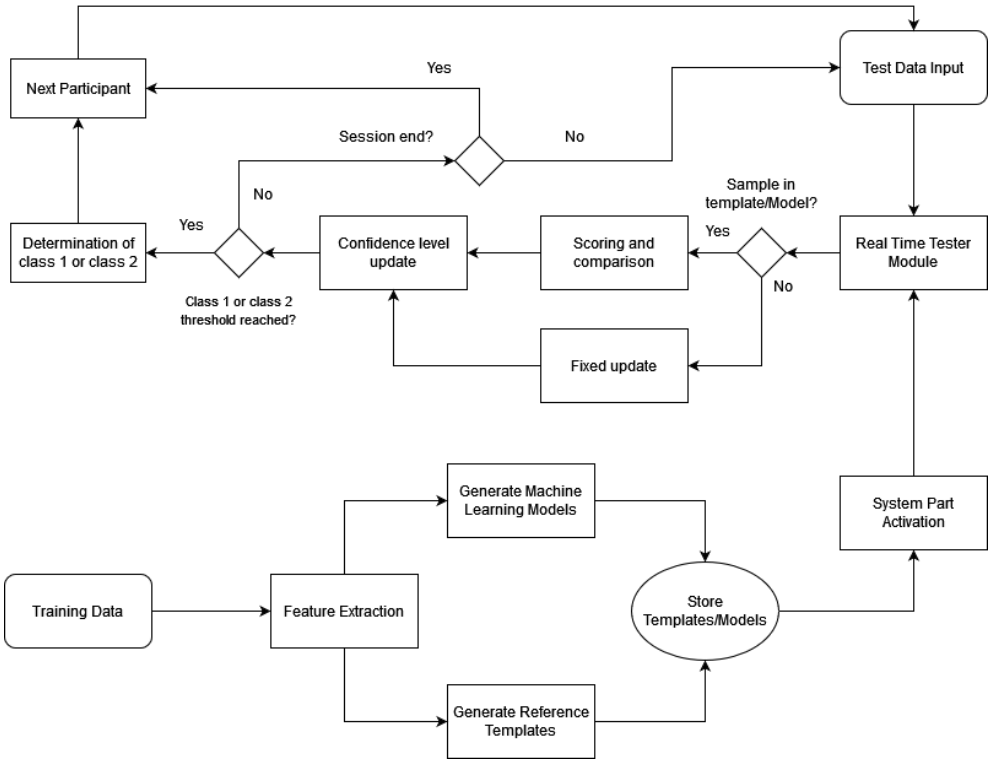


Figure 5.7: System Architecture

biometric models based on the biometric data, and stores them for testing. Furthermore, parts of the system can be activated or deactivated to allow for testing solely statistical, or different machine learning approaches as well as combining these. The system will test based on a participant by participant basis, and create graphs and statistics based on this process. The system will also test differing thresholds to find the best threshold for the confidence determining the gender and age as early as possible. To build the biometric templates focused around one-to-one comparison for the different classes, including different statistical methods discussed prior will be implemented and tested. This will allow for updates being made on singular keystrokes as well in combination with the fixed update if there is test samples missing from the test data input part of the system. Furthermore, to allow for possible machine learning methods, the generation of machine learning

[2] https://www.python.org/

biometric models is also included in the system for both vector-based and singular input prediction. An important aspect to remember here is that vector-based prediction will need to wait for input to predict and will not react for every sample. For instance, a score derived from 10 words will result in the need to wait for these 10 words. However, as samples are presented to the system from test data, the system will react and may as a result, update the confidence level.

5.7.1 Confidence Level

The confidence level is similar to the gender level progression seen in [2] where the aim was early gender detection utilizing messages and conversations. The confidence level will be used to continuously determine gender and age early. The confidence will be set to a static value 0.5. This is because this binary classification problem focuses on males as 1 and females as 0. When the confidence is moving towards 0, the system becomes more and more confident that the participant is female. When the confidence moves towards 1, the system becomes more confident that the participant is male. This is also true for the age classes. The same process for gender will be applied to age. However, there will be thresholds present in the system. The threshold is the point where the system's confidence may or may not reach; thereafter, the system determines if the participant belongs in one of the classes. This is to see the difference in keystrokes needed and the system's accuracy when presenting different thresholds, in other words, how quickly or slowly the system determines depending on the configurations. When there are figures present representing the confidence levels, the blue line is for class label 1, and the red line is for class label 0.

The system or methods can change throughout the experimentation. Therefore, when this occurs, it will be specifically mentioned throughout. The ML models will be trained on the DurationA and the RPlat using the process defined in Section 5.7.3. This is to allow for a different modality than what is used otherwise. In turn, the machine learning part will update the confidence level for every combination of two keys that are present in the machine learning models, which is the same amount of keystrokes necessary as the D2 distance measure discussed in Section 5.7.2. The confidence level will move up or down depending on the scores derived from the methods discussed in the following sections.

5.7.2 Statistical Methods

The scaled Manhattan distance is widely used within the domain of keystroke dynamics in terms of statistical methods. For the following computations, K is for the sample keystroke, T is the biometric template, and i is the instance. Mean after T is the mean for the current biometric template reference key, and std is the standard deviation for the current biometric template reference key. This will be applied to durations with the following computation $SMD = |K_i - Tmean_i| / Tstd_i$. The result of these calculations will provide the distance from the test input toward the reference biometric template of the respective classes. The SMD will also be

used for calculating distance for RPlat to see how well it performs for latency as well. Also, there will be another distance measure for two key combinations from [5] which will be defined as:

$$d_{dur,p} = \frac{|\mu_{dur,p} - t_{dur,p}|}{\sigma_{dur,p}}$$
$$d_{dur,q} = \frac{|\mu_{dur,q} - t_{dur,q}|}{\sigma_{dur,q}}$$
$$d_{lat,pq} = \frac{|\mu_{lat,pq} - t_{lat,pq}|}{\sigma_{lat,pq}}$$

$$D2 = \frac{(d_{dur,p} + d_{dur,q} + d_{lat,pq})}{3}$$

Following this, it is necessary to determine whether there will be an update for classes 1 or 2. For this, the following computation is performed: $D = d_{class1} - d_{class2}$. Then if $d_{class1} > d_{class2}$, the score will move negatively, class2 in this case is the class label 0 and class1 is the class label 1. Otherwise, D is positive. Furthermore, with confidence level as C then $C + (D/100)$. The distance will be divided by 100 to reduce the effect the score has on the confidence level as the system will be reacting for every sample present in the reference template.

5.7.3 Machine Learning

The purpose of applying machine learning is to allow for other modalities than solely distance measures used on durations. The modality for single keystrokes is weaker than multiple. Therefore, it is necessary to consider furthering the accuracy and reducing the keystrokes needed by applying a modality with more information. Therefore, machine learning will be applied in the system, and a grid search from the GridSearchCV and the RandomizedSearchCV module from the sklearn.biometric model_selection library[3] [4]. These were used to automate the process of tuning hyperparameters towards retrieving the best configurations according to the test. The biometric features used are the Duration, including the RPlat. Furthermore, the weights w1, w2, and w3 are determined based on the following, mlx being the certainty of the prediction for the individual ML:

```
mlx <= 0.50                    w1 = 0
mlx > 0.50 and mlx <= 0.75     w1 = 0.5
mlx > 0.75 and mlx <= 0.87     w1 = 1
mlx < 0.87 and mlx <= 1        w1 = 2
```

As a result of this, if the certainty of the machine learning method if less or equal to 0.50 the system will not do anything with the resulting value. The input from the machine learning method will be nullified. It is also necessary to determine whether the following score is positive or negative.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.biometric model_selection.GridSearchCV.html

[4] https://scikit-learn.org/stable/modules/generated/sklearn.biometric model_selection.RandomizedSearchCV.html

```
if mlx predicts 0:
    mlx = -mlx
else:
    mlx = mlx
```

The following is then computed to create a score: $Score = (w1 * ml1 + w2 * ml2 + w3 * ml3)/(w1 + w2 + w3)$. This allows for the biometric models to provide their input on the gender and age classification, followed by computing the average of the three modalities. Furthermore, this result was divided by 100 as the system will use this score when the appearance of a key combination of 2 keys is present and update the confidence positively or negatively depending on the foregoing prediction. With the confidence level as C then $C + (Score/100)$. Furthermore, the data balancing methods will randomly under-sample the training data if one class has more data than the other for training the biometric models. This is towards combating overfitting, which may make the machine learning method biased towards one class, resulting in high specificity or sensitivity recall. The machine learning algorithms chosen were SVM, KNN, and RF as these performed well in terms of state of the art, and were also used in early detection of gender [2], which will now be applied in a continuous fashion for both age and gender. These were implemented by utilizing the sci-kit [5] library in Python.

5.7.4 Fixed Score

Depending on if the fixed score is activated or deactivated within the developed system, the system can react on keystrokes that are not present in the biometric template based on prior updates. Therefore, if this part of the system is activated, and the last confidence update is positive, the fixed update will provide a minor positive update to the confidence. This value is set to 0.00025 because higher than this will make big updates to the confidence level, as the confidence level value is at the lowest 0 and at the highest 1.

5.8 Performance Metrics

Performance metrics for the system are needed. This is to understand how good or poor the system performance is throughout the tests. As this study is dealing with binary classification. The metrics library from scikit[6][7][8] was utilized. The higher the precision, the more true positives the system is able to predict. The lower the precision, the more false positives the system is predicting. This is used to determine if the system predicts that a participant is male, and they are in truth male. Sensitivity is then calculated as $Sensitivity = TP/TP + FN$, whilst the specificity

[5] https://scikit-learn.org
[6] https://scikit-learn.org/stable/modules/biometric model_evaluation.html
[7] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html
[8] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html

is calculated as $Specificity = TN/TN + FP$. The specificity is calculated to give information on how well the system can predict the negative class. In this study case, this will be the female class which is 0, and the younger than 27 class which is also 0. The male participants and the participants of age 27 and above is class 1 and will therefore information regarding their prediction will come from the sensitivity. Considering that an important metrics of this study is to determine the accuracy, the scikit[9] library that computes accuracy scores will be used. As this is a binary classification problem of two classes for both the gender and age classes respectively, this will in turn compare the predicted and true class labels and in turn give an accuracy score in decimals.

There will be tables present to present the results achieved from the tests. These tables will contain the foregoing metrics. Furthermore, the methods used will also be present in the columns. The mean keystrokes needed will also be present in the tables. From these metrics, it will be possible to determine how well the system performs. Regarding state of the art and the varying results seen in literature, a satisfactory compromise of determination speed and accuracy will be approximately 500 keystrokes and 60-70% accuracy for age, as the classes are close without the gap between age classes as seen in literature [3]. Whilst 200-300 keystrokes and higher than 70% accuracy for gender, considering the results from literature on gender detection [2]. However, the goal is to move lower in terms of determination speed whilst attempting to keep accuracy.

[9]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

Chapter 6

Results

This chapter presents the most significant part of most research work. It will summarize the primary findings, including the contributions of the performed study. Furthermore, this chapter will further discuss the implications this research has for the research area, weaknesses, and future work. This chapter also aims to motivate further research and discussion within the field and contribute towards advancing knowledge in the subject matter. However, at this point, the data has been analyzed and processed, and the system will be tested.

6.1 Continuous Determination

This section will provide the results gained when testing the different methods towards determining gender continuously. CKD of gender and age is not as well researched as other determination methods of KD. Therefore, there is trial and error to reach a result that will provide satisfactory metrics. This section will consist of early testing and discussion.

6.1.1 Shuffled Test - Gender

The test was performed on a part of the dataset with the data shuffled. The results can be seen in Table 6.1 is the result from the shuffled test. The threshold present in the tables refers to the confidence threshold that is set. As can be seen in the test result, the higher the threshold, the longer the system takes to determine but it can result in higher accuracy. In total, there will be the possibility of determining 100 participants that have been shuffled. For the test in Table 6.1, the test results are satisfactory with a good mean accuracy all around with the best being 70% accuracy for a mean of 393 keystrokes. The sensitivity recall and specificity are also satisfactory. This indeed displays a balanced update mechanism for the confidence in the Test ID 1 as can be seen in Table 6.1. Without the fixed update mechanism in place, the settings in Test ID 1 is a good compromise of keystrokes needed and accuracy.

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity	Specificity
1	0.4-0.6	Durations	SMD	393	0.70	0.68	0.71
2	0.3-0.7	Durations	SMD	574	0.75	0.73	0.76
3	0.2-0.8	Durations	SMD	828	0.83	0.92	0.77
4	0.1-0.9	Durations	SMD	1010	0.86	1.0	0.74
5	0.0-1.0	Durations	SMD	1194	0.87	1.0	0.73

Table 6.1: Gender with Thresholds - No fixed Update

Furthermore, the keystrokes needed become higher. The tests display a lower sensitivity, indicating that the system is worse at determining if the participants in the test are male versus female.

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity	Specificity
1	0.4-0.6	Durations	SMD + Fixed	517	0.69	0.50	0.88
2	0.3-0.7	Durations	SMD + Fixed	756	0.71	0.5	0.97
3	0.2-0.8	Durations	SMD + Fixed	912	0.75	0.53	1.0
4	0.1-0.9	Durations	SMD + Fixed	998	0.84	0.64	1.0
5	0.0-1.0	Durations	SMD + Fixed	1251	0.95	0.86	1.0

Table 6.2: Gender with Thresholds - With Fixed Update

However, with all things considered, the keystrokes needed are not quite low enough, however, keeping a mean accuracy of 69% with the highest mean accuracy of 95%. Although 95% accuracy is very good, the system does make a decision on all the participants. In fact, at Test ID 5 in Table 6.4 the system only makes a decision on 45% of the participants in the test data set.

However, a test on a shuffled part of the participants in the data set is insufficient. A discovery throughout was that there are participants that make correct decisions from the system difficult. These will be more present in the leave-one-out tests performed in the following section.

6.2 Leave One Out Tests

The following test will be performed on each participant in the dataset once to see how the system performs when testing for each of the 56 participants.

6.2.1 SMD Test - Gender

The SMD will be applied to test. By applying the SMD and considering the participants that fell into their opposite classes, the accuracy decreased. However, there was an increase in keystrokes needed for the system to determine. As seen in Table 6.3, the accuracy even starts to deplete at Test ID 4 and Test ID 5. This is because the ones that were determined correctly in the last test were not determined at all because the threshold was to high. Also, the system has low specificity towards the highest thresholds, as a result of this there are few males that are predicted. The system then has an easier time with predicting the female class. The confidence updates can be seen in Figure 6.1. As seen, the system is quite unsure regarding

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity	Specificity
1	0.4-0.6	Durations	SMD	448	0.72	0.77	0.65
2	0.3-0.7	Durations	SMD	804	0.75	0.86	0.63
3	0.2-0.8	Durations	SMD	921	0.80	0.94	0.64
4	0.1-0.9	Durations	SMD	1098	0.75	0.93	0.53
5	0.0-1.0	Durations	SMD	1151	0.70	0.92	0.28

Table 6.3: SMD Gender with Thresholds

one of the participants, as it used over 1400 keystrokes to determine the gender, while also wrongly classifying a female as male within this range. However, most of the determination is done in the 200-600 keystroke range. The most accurate mean keys necessary are at the threshold 0.2 - 0.8. However, to be able to determine as early as possible, the threshold at 0.4 - 0.6 has an acceptable accuracy of 72%. The reason being is that the determination speed is much faster.

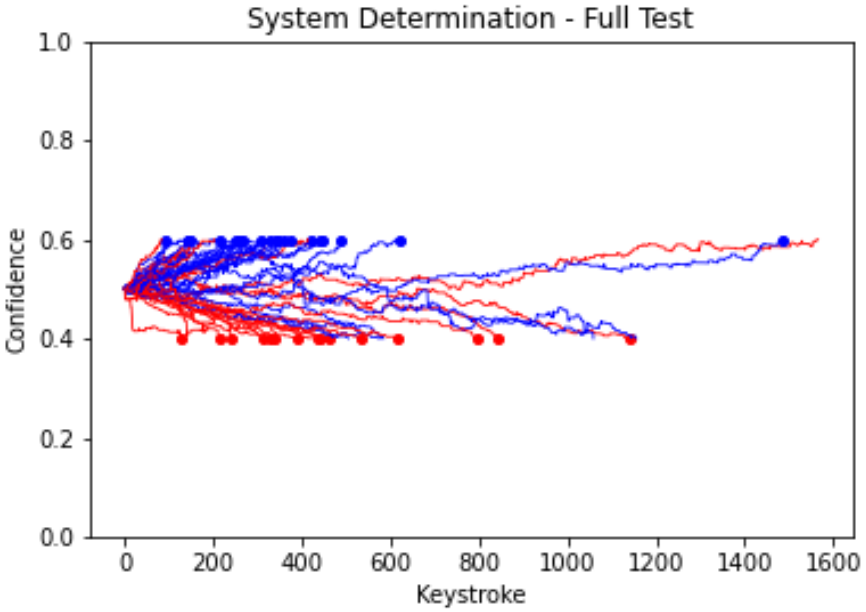


Figure 6.1: Table 6.3 - Test ID 1

The following tests will be done to see how well the SMD is utilized including more distinct features in terms of RPlat . The accuracy was quite stable around 77% which is good. The accuracy is 70% towards 80% for the RPlat . The best result can be seen in Table 6.4 for Test ID 1 with the lowest amount of mean keystrokes of 686 and 77% accuracy. Also, at the highest thresholds, the specificity was 0. This is because the female class did not reach the threshold at such a high threshold, and, therefore, the system did not classify them.

This can be further seen in Figure 6.2, the updates are generally larger than when using only duration's. Something interesting can be seen, which is the same

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity	Specificity
1	0.4-0.6	RPlat	SMD	686	0.77	0.82	0.70
2	0.3-0.7	RPlat	SMD	1047	0.78	0.81	0.71
3	0.2-0.8	RPlat	SMD	1230	0.78	0.88	0.63
4	0.1-0.9	RPlat	SMD	998	0.79	0.93	0.50
5	0.0-1.0	RPlat	SMD	1251	0.77	0.93	0.0

Table 6.4: Gender with Thresholds - RP

trend as in the duration tests. By looking at the male participant in Figure 6.2 follows a similar pattern as the male participant in Figure 6.1 even with different features.

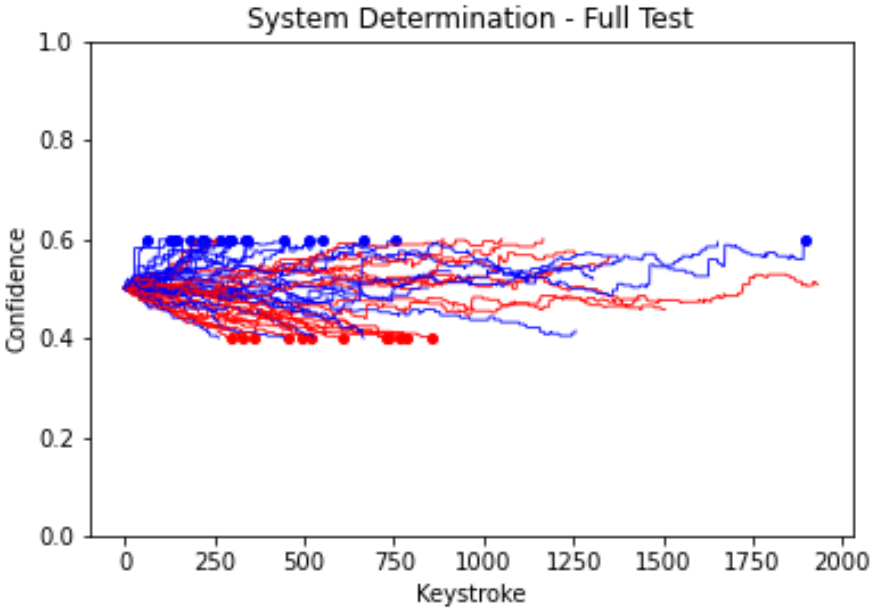


Figure 6.2: Confidence - SMD Gender RP

6.2.2 SMD and Fixed Update - Gender

By introducing the fixed update, it was envisioned that with the SMD the keystrokes needed to determine would be drastically reduced. This was the case, however, not for Test Id 1 as can be seen in Table 6.5. However, for Test ID 2 there are a much lower keystrokes needed at the cost of 5% of accuracy, which is indeed good. Also, the accuracy is continuously increasing. The best accuracy for the test in Table 6.5 is the mean accuracy of 74% and the mean keystrokes of 807, in turn being the threshold 0.2-0.8. This resulted in worse performance than the prior test. Test ID 2 from Table 6.5 can be seen in Figure 6.3. What can be seen here versus the test in Figure 6.1 is that the confidence updates form a

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity	Specificity
1	0.4-0.6	Durations	SMD + Fixed	420	0.67	0.70	0.65
2	0.3-0.7	Durations	SMD + Fixed	573	0.70	0.72	0.68
3	0.2-0.8	Durations	SMD + Fixed	807	0.74	0.77	0.71
4	0.1-0.9	Durations	SMD + Fixed	1038	0.73	0.70	0.68
5	0.0-1.0	Durations	SMD + Fixed	1123	0.80	0.93	0.66

Table 6.5: SMD Gender with Thresholds - Fixed

more continuous line towards the confidence thresholds. The best result for this test is a mean accuracy of 74% and a mean keystrokes of 804, in turn being the threshold 0.2-0.8. This resulted in worse performance than the results Table 6.1.

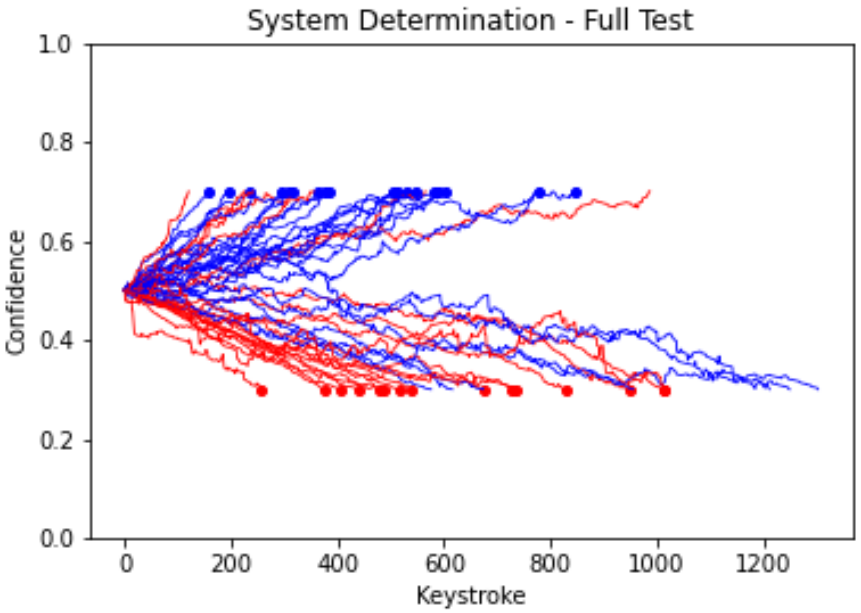


Figure 6.3: Table 6.5 - Test ID 2

Considering that the base system works to a satisfactory manner, the machine learning methods will be applied to see the impact on the determination.

6.2.3 Machine Learning Test - Gender

Following this, the machine learning part of the system is tested. The machine learning part was trained on Duration and RPlat as this is what has been used throughout. The accuracy is in the thresholds 0.4-0.6 at 73% as seen in Table 6.6. However, the keystrokes necessary seem higher than in prior tests but not by a large amount. The system is also quite balanced, which can be seen in sensitivity and specificity. However, the accuracy seems to degrade when the thresholds increase. As seen in the sensitivity recalls, the amount of determined participants becomes less and less. The best accuracy is at 834 keystrokes and 73% accuracy.

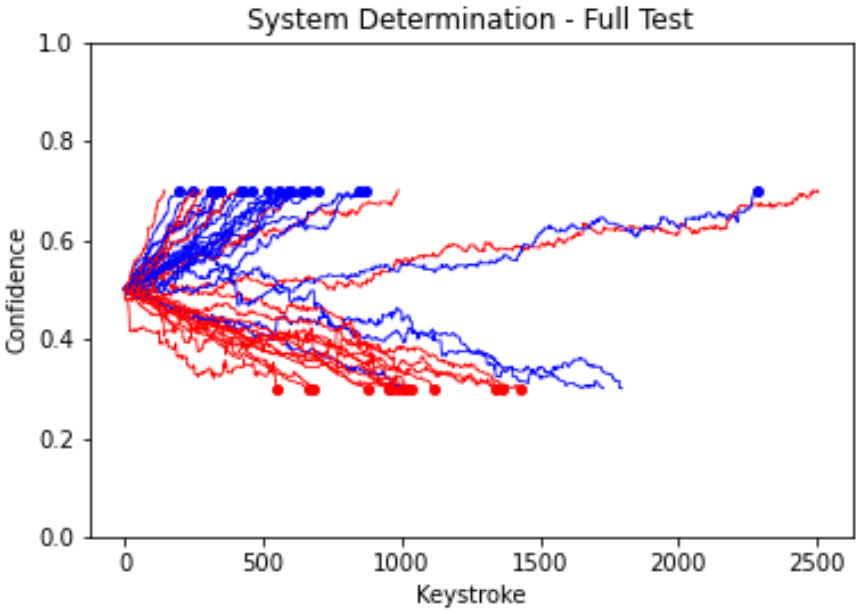


Figure 6.4: Table 6.3 - Test ID 2

This is because the keystrokes necessary to determine with 92% accuracy could equally well benefit more from a method that focuses on periodic keystroke dynamics instead of single samples. This is because it is quite the amount of keystrokes necessary.

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity	Specificity
1	0.4-0.6	Durations + RPlat	ML	843	0.73	0.71	0.75
2	0.3-0.7	Durations + RPlat	ML	1229	0.70	0.68	0.75
3	0.2-0.8	Durations + RPlat	ML	1386	0.68	0.71	0.63
4	0.1-0.9	Durations + RPlat	ML	1630	0.92	1.0	0.83
5	0.0-1.0	Durations + RPlat	ML	1251	0.50	0.0	0.66

Table 6.6: Gender ML with Thresholds

At the threshold 0.1-0.9 the accuracy is 92%. This is indeed quite high. However, prediction is performed on 23% of the participants. At the thresholds of 0.0-1.0, the system loses most participants and has terrible results because it loses the correctly classified participants whilst keeping the ones who are not conforming to the models. The participants that do not conform to the model persist even at this threshold.

6.2.4 Combined Test - Gender

As can be seen in Table 6.7, the system became more stable. This can be seen at the sensitivity recall and specificity and their development, as well as the F1 score. Also, the keys needed to determine has been drastically reduced when combin-

ing the methods. This is due to the faster confidence updates when there are 3 different update functions reacting on the test input.

The accuracy is lower than the other tests, resulting in 3% and 4% accuracy loss for the 0.4-0.6 threshold than when only duration or ML was used. This is in turn, not a big accuracy loss. The best performance gained by combining all the methods were at threshold 0.4-0.6 with an accuracy of 69% with mean keystrokes necessary of 276.

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity Recall	Specificity
1	0.4-0.6	Durations/RPlat	SMD + ML	276	0.69	0.67	0.70
2	0.3-0.7	Durations/RPlat	SMD + ML	429	0.69	0.70	0.68
3	0.2-0.8	Durations/RPlat	SMD + ML	579	0.68	0.68	0.70
4	0.1-0.9	Durations/RPlat	SMD + ML	755	0.68	0.66	0.70
5	0.0-1.0	Durations/RPlat	SMD + ML	859	0.70	0.70	0.70

Table 6.7: Combined - Gender with Thresholds

The most ideal test from Table 6.7 can be seen in Figure 6.5. However, as can be seen, the male participant wiggles a lot between the classes and reach the decision threshold at 1600 keystrokes at one of the lower thresholds displaying that the system may take a long time to determine particular participants even when multiple methods are applied together.

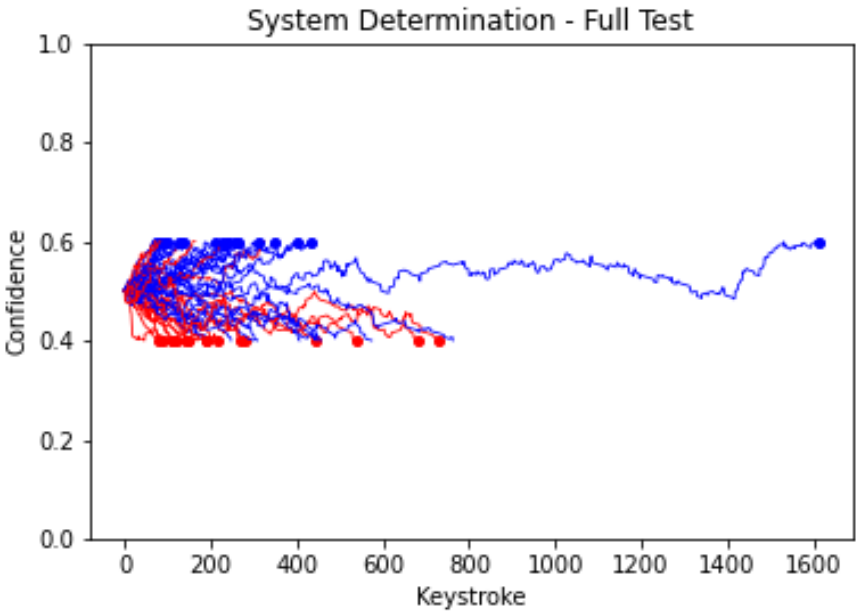


Figure 6.5: Table 6.7 - Test ID 1

6.2.5 Shuffled Test - Age

The following tests will be performed by shuffling the test and training data. As seen in Table 6.9, the system is deciding towards a high amount of the older adult class. Thereafter, the sensitivity recall becomes high and the specificity is quite low. The best result for this test is with 645 mean keystrokes and a mean accuracy of 69%.

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity Recall	Specificity
1	0.4-0.6	Durations	SMD	517	0.55	0.68	0.38
2	0.3-0.7	Durations	SMD	645	0.69	0.94	0.37
3	0.2-0.8	Durations	SMD	884	0.71	0.97	0.36
4	0.1-0.9	Durations	SMD	1074	0.69	1.0	0.0
5	0.0-1.0	Durations	SMD	1285	0.68	1.0	0.0

Table 6.8: Age with thresholds - No fixed update

However, by applying the fixed update function for the shuffled test with age, the accuracy became higher, and the keystrokes necessary before determining the gender became lower as seen in Table 6.9.

This is a similar trend as seen in the gender tests. However, the age determination for these tests responded positively to the fixed update. The system also became more balanced, as seen in the sensitivity and specificity similar to the test in Table 6.5.

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity	Specificity
1	0.4-0.6	Durations	SMD	490	0.68	0.62	0.73
2	0.3-0.7	Durations	SMD	791	0.71	0.62	0.83
3	0.2-0.8	Durations	SMD	884	0.85	0.78	0.92
4	0.1-0.9	Durations	SMD	1305	0.84	0.78	0.91
5	0.0-1.0	Durations	SMD	1498	0.87	0.88	0.86

Table 6.9: Age with thresholds - With Fixed Update

The best result for this test was Test ID 3 in Table 6.9 with mean keystrokes of 884 and an accuracy of 85%. This is due to the high increase in accuracy. When the confidence updates are correct, the fixed update is working well. However, when the confidence updates move towards the opposite classification, the fixed update will push the system toward determining the wrong class.

When applying all the methods for these tests, the system determines with high accuracy as seen in Table 6.10. The test with Test ID 5 displays that the ones in which a decision has been reached conform entirely to the templates and models created.

However, this is the same problem with the shuffled test for gender. This is in regards to the participants that are outliers in terms of their classification. In the following section, these will be included and greatly impact the accuracy of the system.

Test ID	Threshold	Type	Method	Mean Accuracy	Mean Keystrokes	Specificity	Sensitivity
1	0.4-0.6	Durations/RPlat	SMD + ML	0.76	504	0.92	0.64
2	0.3-0.7	Durations/RPlat	SMD + ML	0.80	793	0.97	0.66
3	0.2-0.8	Durations/RPlat	SMD + ML	0.96	1146	1.0	0.92
4	0.1-0.9	Durations/RPlat	SMD + ML	0.97	1387	1.0	0.95
5	0.0-1.0	Durations/RPlat	SMD + ML	1.0	1565	1.0	1.0

Table 6.10: Age with Thresholds

6.2.6 Leave One Out Tests - Age

As can be seen when using the SMD in its singularity in Table 6.11, the best result from the test is Test ID 4. This is not a good compromise of accuracy and keystrokes needed, as the system uses too long to determine the age.

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity Recall	Specificity
1	0.4-0.6	Durations	SMD	416	0.56	0.56	0.57
2	0.3-0.7	Durations	SMD	655	0.57	0.66	0.47
3	0.2-0.8	Durations	SMD	972	0.61	0.73	0.47
4	0.1-0.9	Durations	SMD	1092	0.71	0.86	0.53
5	0.0-1.0	Durations	SMD	1189	0.76	1.0	0.44

Table 6.11: SMD Age with Thresholds - No Fixed

As seen in Figure 6.6, the system struggles with separating the two age classes. This is because it is difficult to separate the classes, as they are quite close. As a result, it is more difficult to separate and classify the participants in their respective age groups.

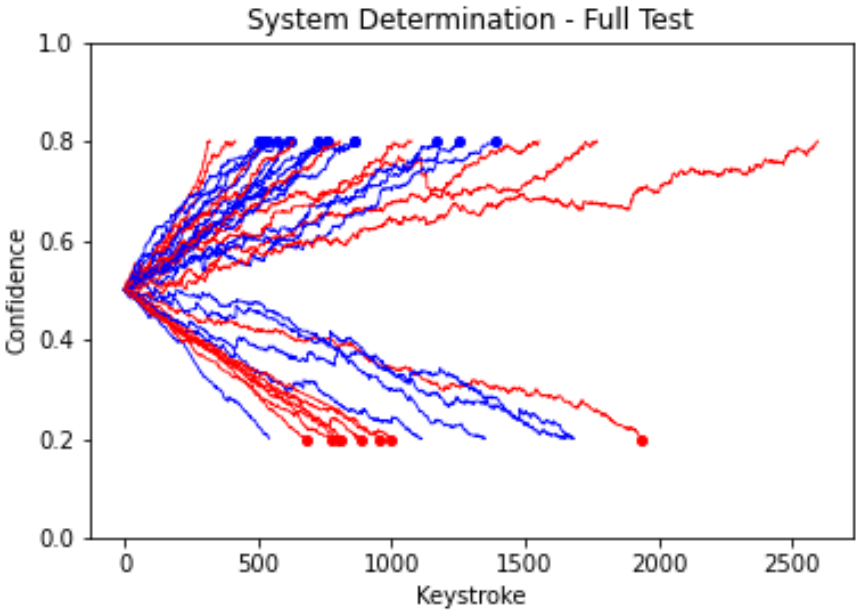


Figure 6.6: Table 6.11 - Test ID 3

It is envisioned that the age determination would benefit from the fixed update function in the system. This was not the case. The system became more balanced in terms of the different classes, which is positive. However, the accuracy went down and there was not much reduction in keystrokes. Therefore, the system needs more accuracy including more frequent confidence updates for the age determination. The best result here was with Test ID 4 in Table 6.12 with 917 keystrokes and 63% accuracy. This is not good results in either keystrokes or accuracy.

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity	Specificity
1	0.4-0.6	Durations, RPlat	SMD	395	0.58	0.51	0.65
2	0.3-0.7	Durations, RPlat	SMD	647	0.56	0.56	0.57
3	0.2-0.8	Durations, RPlat	SMD	772	0.59	0.66	0.50
4	0.1-0.9	Durations, RPlat	SMD	917	0.63	0.68	0.57
5	0.0-1.0	Durations, RPlat	SMD	1189	0.65	0.68	0.61

Table 6.12: SMD Age with Thresholds - Fixed

6.2.7 Combined Test - Age

A test with all the methods applied will be tested to see what impact this has on the age test as seen in Table 6.23. The age test seems to not improve much in terms of accuracy in contrast to the fixed update. Other than that the keys needed to determine has been lowered, but none of the tests display satisfactory accuracy.

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity Recall	Specificity
1	0.4-0.6	Durations, RPlat	SMD + ML	334	0.57	0.51	0.62
2	0.3-0.7	Durations, RPlat	SMD + ML	601	0.54	0.53	0.55
3	0.2-0.8	Durations, RPlat	SMD + ML	707	0.56	0.58	0.52
4	0.1-0.9	Durations, RPlat	SMD + ML	905	0.61	0.66	0.55
5	0.0-1.0	Durations, RPlat	SMD + ML	929	0.65	0.70	0.60

Table 6.13: Combined - Age with Thresholds

6.3 New System Test

During the prior testing, new functions were implemented in the system in an attempt to handle the participants that become outliers in terms of their class. This seems to be the biggest challenge, as the system gains high accuracy and a satisfactory low amount of keystrokes depending on the system configuration and features chosen. Therefore, it displays the possibility of increasing the accuracy when managing this factor when dealing with difficult data sets such as but not limited to free text.

Therefore, the system has been updated with the following. The templates for the duration is now divided into one with the most discriminating features and one for the most used features. This is to allow for different weighting of the two templates and as a result, allow for higher confidence updates for the more

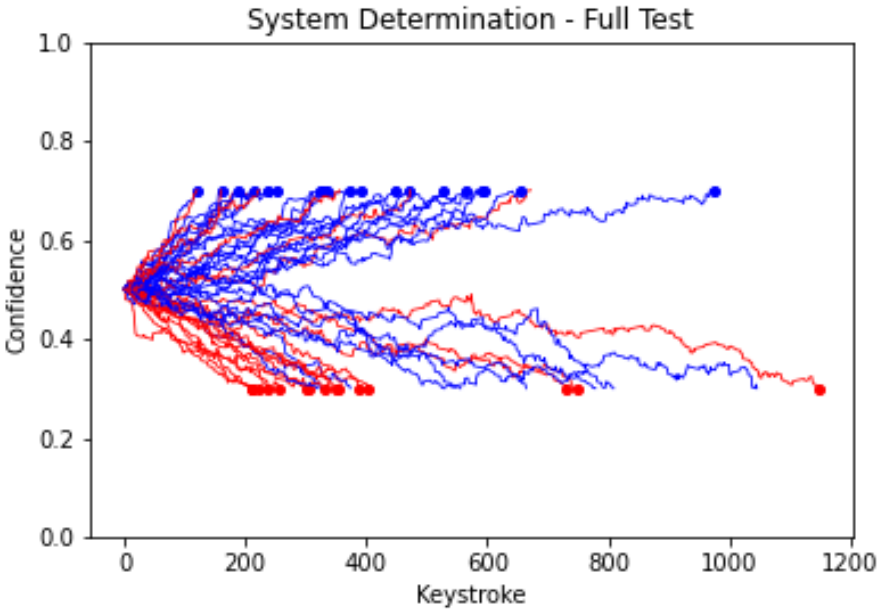


Figure 6.7: Updated System

important durations. An alert system where after a certain amount of keystrokes towards one class, the test participant will gain higher confidence updates moving forward. This allowed the system to move up or down depending on the behavior over time. The following tests will consist of a full test of gender to see if the full update of the system provides better results. The configurations were the same as in Table 6.7 except for introducing the new functions to the system. Upon testing the system as seen in Figure 6.7, the system managed an accuracy of 70%, which is similar to Test ID 2 in Table 6.7. However, the system is reacting more to samples. However, the test resulted in a mean keystroke necessary to determine 429 keystrokes, the same amount as the prior systems test. This method did not work well, as the only part that was altered was the frequency of the updates; in other words, the update would move quicker or slower but still in the same direction. Therefore, this did not affect the determination speed or the accuracy in a positive manner. The method might benefit from weighting for aspects of importance rather than alterations in divisions. Because of the time constraints of the study, there were no further efforts towards this part of the system, and the system was reset to its original state for the following tests.

6.4 KLD and Probability Implementation

As it is now know that there are less features to chose from in terms of the age groups, and that it needs to include features that are have more predictive power.

The following tests will apply the KLD and probability feature selection method discussed in Section 5.5.1. Thereby, introducing biometric features with a larger KLD value instead of a manual visual selection of features utilizing vizualisation of the pdfs. When these methods are applied, they will be specifically mentioned in the tables and figures. The features used will be revised, and new tests will be performed to see how effective it is towards improving accuracy and reducing the keystrokes necessary to gain a satisfactory compromise. As can be seen the test in Table 6.14, the accuracy increased. However, as a result, the system is struggling to determine one of the classes as the specificity after a while becomes 0. This is due to the difficulty of feature choice, as few features can be used to make the system more balanced and efficient for durations. As a similar trend to the gender tests, when including mostly such features, the balance and accuracy degrade as we lose the participants that were classified in the earlier thresholds because the amount of keystrokes contributed by participants vary.

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity	Specificity
1	0.4-0.6	Durations	SMD	1103	0.71	1.0	0.33
2	0.3-0.7	Durations	SMD	1401	0.77	1.0	0.0
3	0.2-0.8	Durations	SMD	1561	0.75	1.0	0.0
4	0.1-0.9	Durations	SMD	1643	0.66	1.0	0.0
5	0.0-1.0	Durations	SMD	1711	0.80	1.0	0.0

Table 6.14: Durations - Higher KLD Durations Age

Because of this, it is difficult making the system efficient in terms of age determination when using durations as the trade-off between accuracy and amount of keystrokes necessary are high including difficult to balance. Although the durations are more often used. More features were distinct for RPlat than in comparison to duration.

The following test was performed by applying features with high KLD value, but low Probability of encountering the features in the dataset. The following results were found in Table 6.15. The accuracy higher as seen in Test ID 1 in Table 6.15. However, the keystrokes necessary increased drastically.

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity Recall	Specificity
1	0.4-0.6	RPlat	SMD	1644	0.875	1.0	0.66
2	0.3-0.7	RPlat	SMD	1717	0.80	1.0	0.50
3	0.2-0.8	RPlat	SMD	1717	0.75	1.0	0.0
4	0.1-0.9	RPlat	SMD	1717	0.75	1.0	0.0
5	0.0-1.0	RPlat	SMD	1722	0.66	1.0	0.0

Table 6.15: RPlat with High KLD Gender

Furthermore, the accuracy drops as the threshold rises. This is because the system is failing to classify the participants that have been correctly classified earlier, and still classifies the ones that has been classified incorrectly in this case. As can be seen in the trend in Table 6.15 and Table 6.3, the female class have less keystrokes. Meaning when the threshold is high enough, there wont be enough keystrokes to classify the female class.

A filtering part was developed for the system, and will allow the system to filter out features depending on the KLD and probability criteria. The following test was performed with SMD for RPlat, including features with the following values: KLD > 1 for RPlat latencies. This was done to see how well this method works. The results gained are present in Table 6.16. These results are much better than the ones received in Table 6.15, although with an accuracy loss of 2.5% the keystrokes necessary for the system to make a decision were reduced with 721 keystrokes for Test ID 1. This indicates that higher KLD indeed gives good accuracy, and the lower the value, the less obscure key combinations will be encountered, resulting in more frequently used features. This will result in the system's ability to make faster decisions, in particularly when utilizing thresholds, as there will be more confidence level updates.

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity	Specificity
1	0.4-0.6	RPlat	SMD	923	0.85	0.90	0.80
2	0.3-0.7	RPlat	SMD	1180	0.86	0.93	0.78
3	0.2-0.8	RPlat	SMD	1361	0.85	0.92	0.75
4	0.1-0.9	RPlat	SMD	1406	0.84	0.92	0.66
5	0.0-1.0	RPlat	SMD	1722	0.83	0.91	0.66

Table 6.16: RPlat with KLD > 1 Gender

A final test will be attempted with KLD > 0.5 for RPlat latencies, to see indeed how low it is possible to go before the accuracy becomes unsatisfactory. As seen in Table 6.17, the accuracy is 79% with a much better keystrokes necessary for Test ID 1. There is, however, a 6% accuracy loss.

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity	Specificity
1	0.4-0.6	RPlat	SMD	512	0.79	0.92	0.63
2	0.3-0.7	RPlat	SMD	678	0.81	0.875	0.69
3	0.2-0.8	RPlat	SMD	700	0.85	0.92	0.55
4	0.1-0.9	RPlat	SMD	784	0.78	0.89	0.55
5	0.0-1.0	RPlat	SMD	928	0.84	0.89	0.66

Table 6.17: RPlat with KLD > 0.5 Gender

By testing with latencies with a probability higher than 0.0010, will include all latencies with an appearance higher than 1%. This is tested to see how the latencies with higher probability with no considerations of the KLD will affect the systems performance. As can be seen in Table 6.18, for the lowest threshold the accuracy is reduced to 64% whilst the keystrokes necessary to make a decision were reduced to 189. This is indeed efficient, however, the accuracy is not satisfactory.

It can also be from the prior test seen in Figure 6.8, that the male participants where a correct decision in that test are now being wrongly classified in the female class in Figure 6.2. This is because there are so many features with low KLD included in the test in Table 6.18 that the updates are becoming more frequent and random.

The machine learning method had the following result seen in Table 6.19. The same trend seen in the prior test is also present for this test in terms of balance

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity	Specificity
1	0.4-0.6	RPlat	SMD	189	0.64	0.625	0.66
2	0.3-0.7	RPlat	SMD	414	0.62	0.61	0.65
3	0.2-0.8	RPlat	SMD	563	0.66	0.63	0.72
4	0.1-0.9	RPlat	SMD	660	0.68	0.65	0.72
5	0.0-1.0	RPlat	SMD	743	0.70	0.70	0.70

Table 6.18: RPlat > 0.0010 Probability Gender

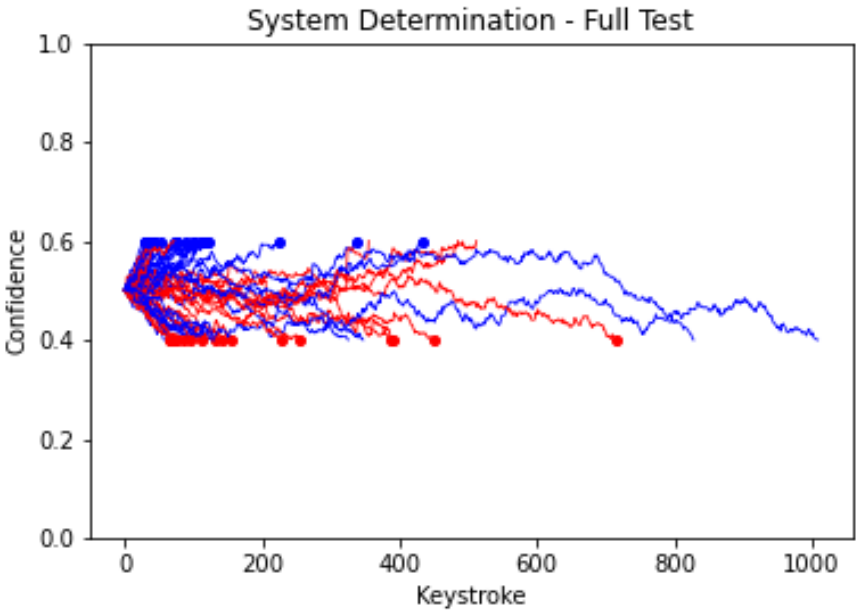


Figure 6.8: High Probability - SMD Gender RP

related to the sensitivity and specificity. This displays that the feature extraction method works for all the system parts.

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity	Specificity
1	0.4-0.6	DuraA + RPlat	ML	843	0.71	0.52	0.91
2	0.3-0.7	DuraA + RPlat	ML	1451	0.69	0.50	0.85
3	0.2-0.8	DuraA + RPlat	ML	1591	0.80	0.85	0.75
4	0.1-0.9	DuraA + RPlat	ML	1726	0.80	0.66	1.0
5	0.0-1.0	DuraA + RPlat	ML	1763	0.50	0.50	0.0

Table 6.19: Gender ML probability >= 0.0010 and KDL >= 0.1

A final test is performed with gender combining the RPlat SMD and ML parts with the fixed update. As seen in Table 6.21, when combining the two prior tests, that were good. The system managed to determine the gender with satisfactory speed and accuracy.

The final result for gender from Table 6.21 can be seen in Figure 6.9, the most efficient test gives 322 keystrokes with 71% accuracy.

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity	Specificity
1	0.4-0.6	DurA + RPlat + DurB	D2	933	0.65	0.42	0.86
2	0.3-0.7	DurA + RPlat + DurB	D2	1595	0.81	0.70	0.91
3	0.2-0.8	DurA + RPlat + DurB	D2	1762	0.85	0.50	1.0
4	0.1-0.9	DurA + RPlat + DurB	D2	0	0.0	0.0	0.0
5	0.0-1.0	DurA + RPlat + DurB	D2	0	0.0	0.0	0.0

Table 6.20: Gender D2 probability >= 0.0010 and KDL >= 0.1

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity	Specificity
1	0.4-0.6	DurA, RPlat	SMD + ML + Fixed	322	0.71	0.65	0.79
2	0.3-0.7	DurA, RPlat	SMD + ML + Fixed	573	0.78	0.75	0.82
3	0.2-0.8	DurA, RPlat	SMD + ML + Fixed	765	0.80	0.78	0.81
4	0.1-0.9	DurA, RPlat	SMD + ML + Fixed	849	0.84	0.84	0.84
5	0.0-1.0	DurA, RPlat	SMD + ML + Fixed	926	0.81	0.83	0.80

Table 6.21: Gender SMD - probability KDL > 0.5 and ML - probability >= 0.0010 and KDL >= 0.1

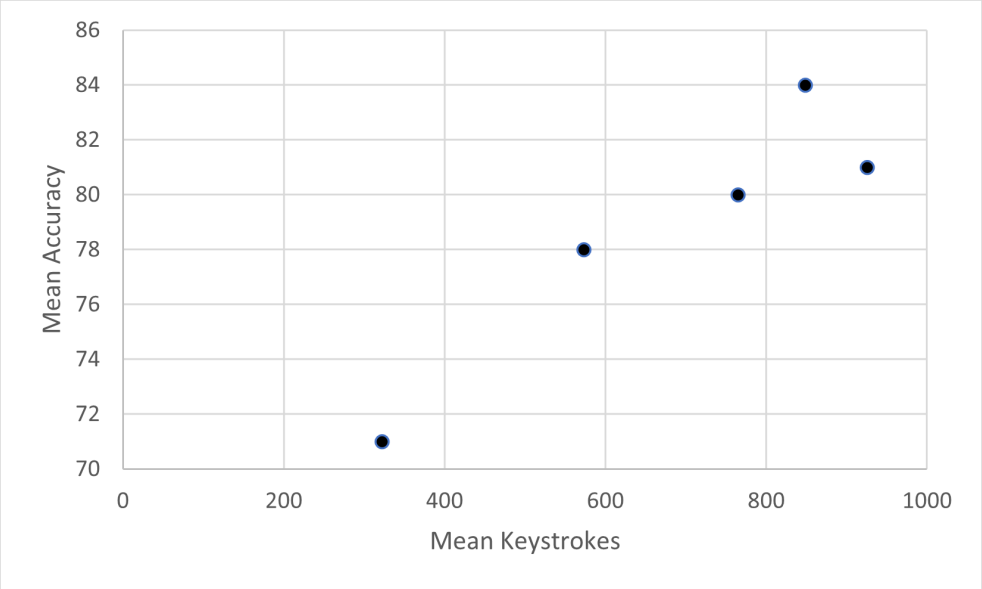


Figure 6.9: Gender - Final Result

There is, however, still one male participant that uses quite an amount of keystroke before the system can make a decision, and can also be seen throughout the other tests and which can still be seen in this test as well in Figure 6.10.

The final test for age will be the same as the final test for gender, just for the age classes instead. As can be seen in Table 6.22. This also worked for age and determined the age after a satisfactory amount of keystrokes with satisfactory accuracy in Test ID 1. As seen with the prior tests on age, it indeed needed more frequent confidence updates whilst including features with higher predictive power.

The final result for the age tests from Table 6.22 can be seen in Figure 6.11, the most efficient test gives 312 keystrokes with an accuracy of 72%.

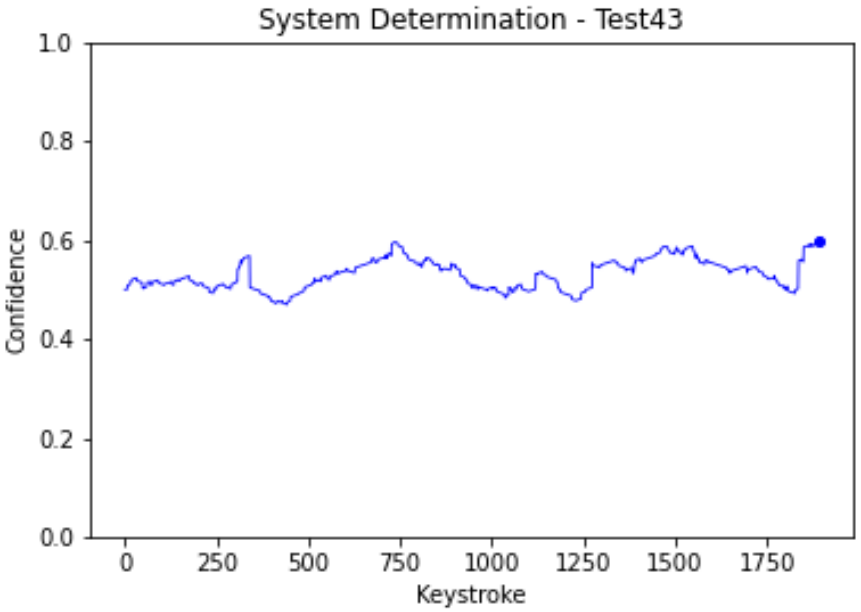


Figure 6.10: Gender - High Amount of Keystrokes

Test ID	Threshold	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity	Specificity
1	0.4-0.6	DurA, RPlat	SMD + ML + Fixed	312	0.72	0.60	0.82
2	0.3-0.7	DurA, RPlat	SMD + ML + Fixed	597	0.69	0.52	0.85
3	0.2-0.8	DurA, RPlat	SMD + ML + Fixed	825	0.77	0.63	0.88
4	0.1-0.9	DurA, RPlat	SMD + ML + Fixed	1054	0.75	0.63	0.86
5	0.0-1.0	DurA, RPlat	SMD + ML + Fixed	1109	0.77	0.66	0.85

Table 6.22: Age SMD - KDL > 0.5 and ML - probability >= 0.0010 and KDL >= 0.1

As can be seen for the age test as well in Figure 6.12, there are also participants that need a longer amount of keystrokes before the system can make a decision. These are quite difficult to manage as they wiggle between the classes.

To conclude the tests, a weakness regarding how this feature selection method was implemented was discovered, and not the method itself. That is, how it was implemented does not consider other features beyond the ones it is used on. For instance, the distance measure D2 that uses the duration from the first key press, the RPlat, then the duration from the second key press. This comparison is not necessarily fair with the SMD for latency. This is because when looking at RPlat, it is known through the feature selection part that this feature indeed has good predictive value. However, the D2 distance may in turn not have a good duration, which in turn can result in poorer results for this distance measure. An example of this is when including the duration of the letter s into the SMD part of the system that uses durations to determine gender, the accuracy will drop drastically because the updates become more or less random as this feature is similar for the gender

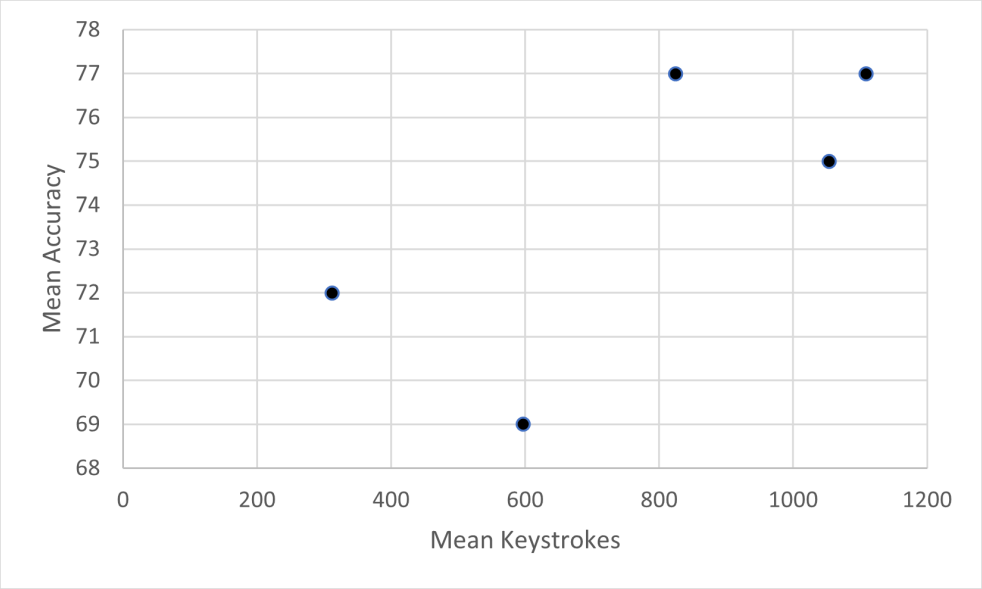


Figure 6.11: Age - Final Result

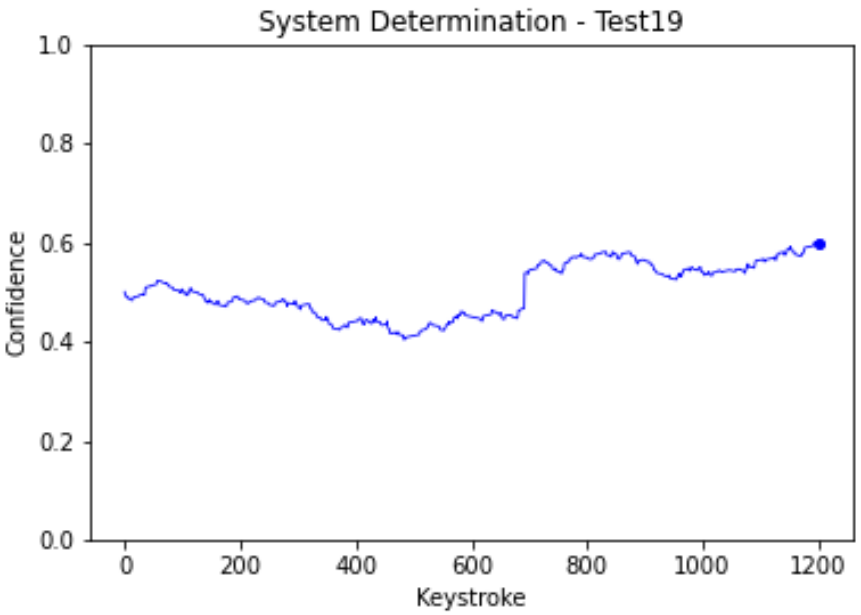


Figure 6.12: Age - High Amount of Keystrokes

classes but frequently used. However, the D2 distance will possibly perform better when this is considered. Considering that the duration for the letter s was part of one of the resulting key combinations used in Table 6.20, it is impressive that the distance measure gained 81% accuracy for Test ID 2 considering that one of

the poorest features was as a result of the foregoing used in the test. Also, the accuracy of the machine learning part of the system started to suffer when there were fewer than approximately 50 instances for class 0 and class 1, respectively. Thereby, 50 instances of a feature for each class were set as a minimum.

6.5 Participant Types

Throughout the tests there were observations in terms of the participant's behavior similar to those described as sheep and goats in Doddingtons zoo menagerie [26]. There were users that had behaviour which affected the systems accuracy, they can be defined as follows. Deviatiors, they deviated to such an extent from their classification within the dataset that the determination was highly affected by them in a negative manner, in turn affecting accuracy greatly. Conformers fall into their classification and in turn, are classified correctly. From these conformers, there are in particular, two types. One type of conformer could quickly be determined within their class; the other slowly moved towards the correct determination. An example of a slow conformer from the gender tests for both females and males can be seen in Figure 6.13 and Figure 6.14. However, an example of a fast fe-

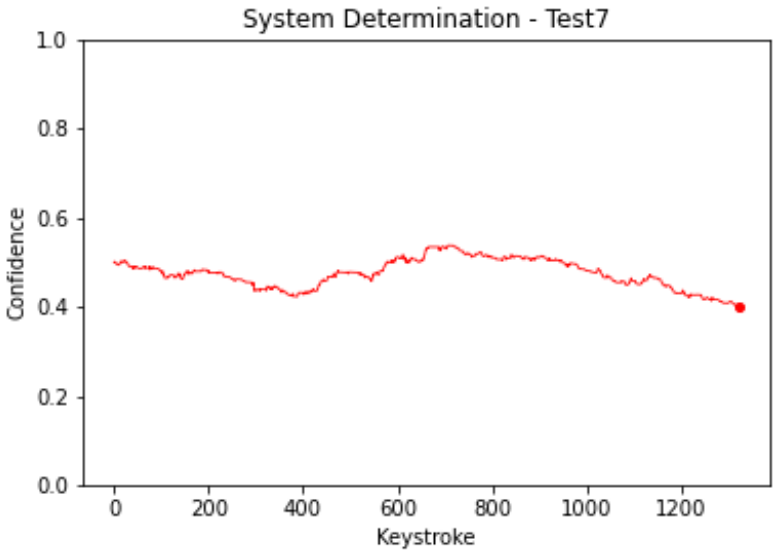


Figure 6.13: Female Conformer

male deviator can be seen in Figure 6.16. This participant acts to such an extent in the other class that the participant is determined wrongly after 100 keystrokes. Another observation from the tests are that generally both male and female participants are belong to these participant types, and the same is true for the age classes also. This indicates that in terms of the data gathering task performed, the typing behaviour for male and female to some extent is predictable, but are more

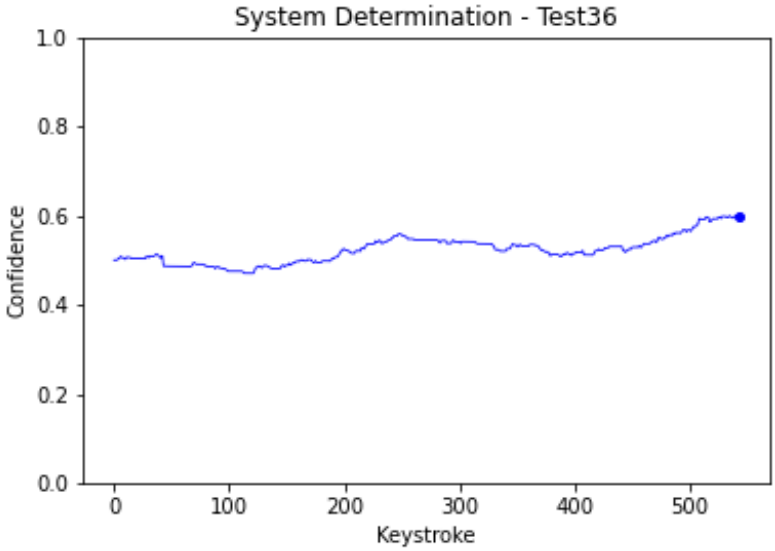


Figure 6.14: Male Conformer

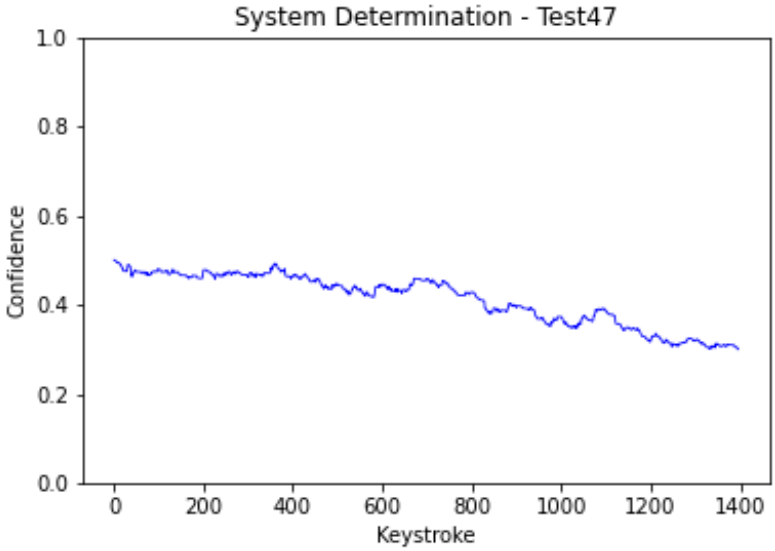


Figure 6.15: Male Deviator

difficult to predict. This is due to nature of the data gathering task including the amount of participants.

The system wiggled between certain and uncertain as seen in Figure 6.13. However, when changing the threshold to 0.2 - 0.8, this participant may not be determined as the confidence never reaches the threshold. Therefore, this participant will become a mediator participant if the threshold is too high. An example

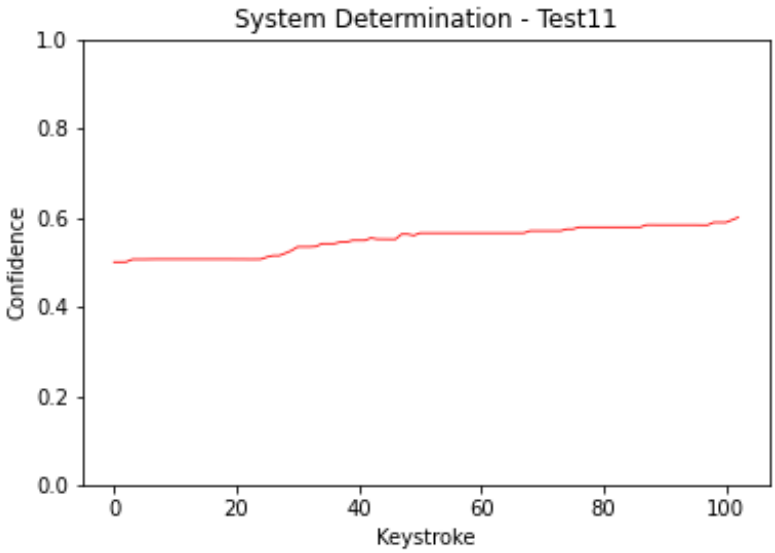


Figure 6.16: Female Deviator

of a male mediator can be seen in Figure 6.17.

Then there are medial participants, which the system could not determine when the thresholds were above a certain value. The challenge was to determine deviators. The deviators were difficult to correctly determine, as most acted largely as the opposite class throughout their typing session. Figure 6.15 shows an example of a slow male deviator. These types of deviators tend to move slowly towards the opposite class, as can be seen.

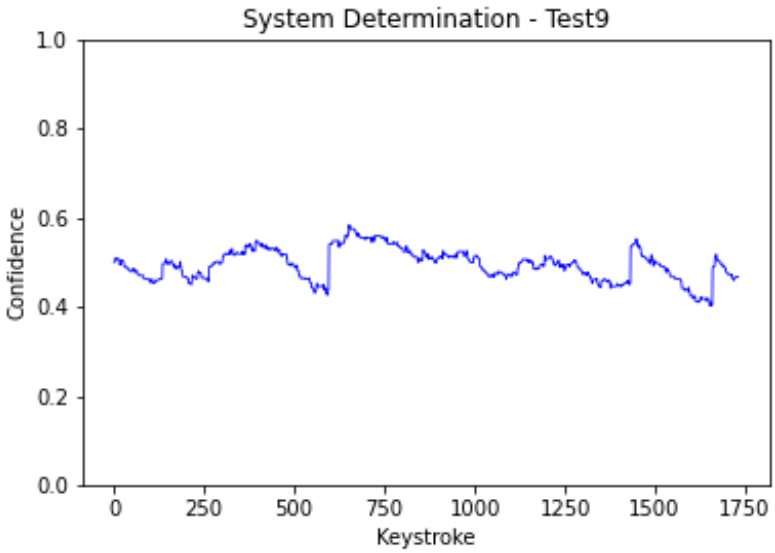


Figure 6.17: Male Mediator

6.6 Voting System

Knowing that the base system is working, and as discussed prior to this, there was a need to handle participants that were outliers. Also, there is always the aim to decrease keystrokes necessary to decide whilst keeping satisfactory accuracy. Therefore, a final method will be applied to gender, including age, to see how the accuracy and keys required will be affected. This method focuses on instead of using thresholds and rather performing voting after an amount of confidence level updates have been performed. Therefore, there will be no graphs of the confidence levels for this section. Because of the limited time available for this study and the workload necessary to complete it, multiple configurations and tests were not performed. However, because of the importance of the possibility of handling outlier participants and reducing the number of keystrokes to gain a satisfactory classification speed, a simpler test was performed. The configurations were as follows. When the voting is completed, the confidence level is reset to its starting point of 0.5. These tests will only include duration as they are single key due to the time constraint, and as a result, only include the gender class. Throughout the test, there were observations where outliers were handled. By including 10 votes, an example of an outlier that was managed correctly can be seen:

```
[0, 1, 0, 0, 1, 1, 1, 1, 1, 1]
```

This participant was male. Therefore, the resulting decision was correct. Furthermore, the participant behaves like the female class for the first, third, and fourth inclusions toward the voting. If the classification had been decided with five votes or earlier, the participant would have been female. Therefore, it is indeed a posit-

ive finding regarding managing outlier participants. Sometimes participants who conform with the template can also be seen, as in the example below with a female participant:

```
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

This indicates that even though this decision function is used, there will still be those who conform and deviates. However, it introduces the possibility of managing them more effectively than when using only thresholds. A small test with good results can be seen when applying for every 10th update with five voting instances with features that has satisfactory predictive power.

Configuration	Type	Method	Mean Keystrokes	Mean Accuracy	Sensitivity	Specificity
10 Updates/5 Votes	Durations, RPlat	SMD/Voting	126	0.71	0.63	0.83

Table 6.23: 10/5 - Gender with Voting

Fewer keystrokes are necessary because the voting system will force the decision before reaching the threshold tested in the preceding section. Furthermore, the accuracy seems to keep itself relatively satisfactory. However, this depends on features that separate the classes early in the typing activities. If the features chosen start to differentiate further in the writing session, but in the early writing stages, they are similar, it can impact the result. Another exciting discussion that can be had regarding this is that the participants may become more used to the task and learn how to use the keyboard more effectively as the typing task moves along. However, this indeed displays further possibilities in the early use of voting in the system.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this study, possibilities towards continuously determining gender and age has been tested. The distance measures that has been tested is the SMD and the D2 distance measure from Section 5.7.2, including a machine learning part of the system from Section 5.7.3, which follows a stepping function with ensemble voting. The SMD performed best on one feature category, both fastest and most accurate, but highly sensitive to feature choice as including poor features greatly impacted the accuracy. The D2 distance measure was more robust because it could include features that were similar in terms of distinctness but still gain high accuracy. The machine learning part of the system worked well but needed approximately 50 instances of each class or else the accuracy would start to drop. The negative side of the D2 distance measure was that it was generally slower than the scaled Manhattan distance for the test; however, it was discovered that this could possibly be due to feature choice. The combination of different system parts to update the confidence level performed well and also reduced the keystrokes necessary. The tests performed do not by any means cover every possible method or configuration towards creating a CKD system to determine age and gender, but rather a limited amount of them and display positive results in that regard. The result with the best compromise was 126 keystrokes and 71% accuracy for gender by utilizing the voting part of the system and 71% accuracy with 322 keystrokes by utilizing the threshold part of the system. The highest accuracy was gained by utilizing SMD on the most distinct features, resulting in 87.5% accuracy with a mean of 1644 necessary to determine gender. For age, the earliest determination gained with satisfactory accuracy was a mean of 312 keystrokes necessary for determination with 72% accuracy. The highest accuracy for determination of age was 77% accuracy with a mean of 825 keystrokes necessary. A biometric feature selection method that suits the criteria in this study for CKD has been chosen, particularly when dealing with a binary classification problem. This resulted in a feature selection method that considered the probability, including the distinctiveness of the respective features. Both the feature selection and statistical methods were

effective in the final tests when combined, including the machine learning part of the system. However, the system may become more efficient with further configurations for machine learning and focused feature selection for the D2 measure, which uses three modalities. The system works towards CKD of gender and age and can combine or exclude the different methods used by activating different parts of the system. The voting part of the system also works; however, not much effort went into testing or configuring this part of the system because of the time limit of the study, and thereby rather displays positive possibilities in that regard and in particular towards managing deviants and further lowering the necessary amount of keystrokes. The system is highly sensitive to feature choice, so much so that choosing features that are similar in terms of distinctiveness and frequently used between classes will result in a large negative impact on the accuracy. If features that were not often used but distinct were chosen, the determination speed would be lowered significantly. The participant's test data session was continuously monitored, and a single and/or a combination of two keystrokes were considered for every sample. Thereafter, the system can continuously determine age and gender. However, possible improvements and future work will be discussed in the following section.

7.2 Future Work

In this study, there have been positive findings regarding the continuous determination of age and gender. However, this domain is vast, thereby, a high amount of future work can be performed in this domain. This section will, however, not discuss all of these possibilities. But rather discuss some of the identified possibilities. The main focus of this study was on durations and RPlat for the classes. It would be interesting to build more specific models and templates; this can be for key combinations on the right side or left side of the keyboard. As seen in Section 5.5.1, the most frequently used combinations have been discussed but not used in their singularity. Furthermore, a more complex feature selection process which includes considerations for class balance and the modalities for the D2 distance measure will possibly increase accuracy and, as a result, allow for other methods to be applied more effectively. Also, as the voting system which resets the confidence level after x amount of confidence level updates indeed displays the potential to handle participants that deviates to some extent from the templates. Further efforts with a method that includes and focuses on this would possibly increase accuracy and reduce keystrokes needed for determination. An example of such a method can be seen in the final test where the voting system is applied, resulting in the fastest determination method for this study.

Bibliography

[1] I. Tsimperidis, A. Arampatzis and A. Karakos, 'Keystroke dynamics features for gender recognition,' *Digital Investigation*, vol. 24, pp. 4–10, Mar. 2018, ISSN: 17422876. DOI: 10.1016/j.diin.2018.01.018.

[2] R. J. Strømme, 'Early gender detection using keystroke dynamics and stylometry,' *Institutt for informasjonssikkerhet og kommunikasjonsteknologi*, 2021.

[3] A. Pentel, 'Predicting user age by keystroke dynamics,' vol. 764, Springer Verlag, 2019, pp. 336–343, ISBN: 9783319911885. DOI: 10.1007/978-3-319-91189-2_33.

[4] O. D. Tverrå, *Determining age with keystroke dynamics*, Biometrics IMT 4126 Report, 2022.

[5] P. Bours, 'Continuous keystroke dynamics: A different perspective towards biometric evaluation,' *Information Security Technical Report*, vol. 17, pp. 36–43, 1-2 Feb. 2012, ISSN: 13634127. DOI: 10.1016/j.istr.2012.02.001.

[6] K. Nilsen and P. Bours, 'Combining periodic and continuous authentication using keystroke dynamics.'

[7] S. Krishnamoorthy, L. Rueda, S. Saad and H. Elmiligi, 'Identification of user behavioral biometrics for authentication using keystroke dynamics and machine learning,' Association for Computing Machinery, May 2018, pp. 50–57, ISBN: 9781450363945. DOI: 10.1145/3230820.3230829.

[8] C. Vesel and et al., 'Effects of mood and aging on keystroke dynamics metadata and their diurnal patterns in a large open-science sample: A bi-affect ios study,' *Journal of the American Medical Informatics Association*, vol. 27, pp. 1007–1018, 7 Jul. 2020, ISSN: 1527974X. DOI: 10.1093/jamia/ocaa057.

[9] A. Pentel, 'Predicting age and gender by keystroke dynamics and mouse patterns,' Association for Computing Machinery, Inc, Jul. 2017, pp. 381–385, ISBN: 9781450350679. DOI: 10.1145/3099023.3099105.

[10] B. Ayotte, M. K. Banavar, D. Hou and S. Schuckers, 'Fast and accurate continuous user authentication by fusion of instance-based, free-text keystroke dynamics; fast and accurate continuous user authentication by fusion of instance-based, free-text keystroke dynamics,' 2019.

[11] A. Alshehri, F. Coenen and D. Bollegala, 'Iterative keystroke continuous authentication: A time series based approach,' *KI - Kunstliche Intelligenz*, vol. 32, pp. 231–243, 4 Nov. 2018. DOI: 10.1007/s13218-018-0526-z.

[12] I. Stylios, A. Skalkos, S. Kokolakis and M. Karyda, 'Bioprivacy: A behavioral biometrics continuous authentication system based on keystroke dynamics and touch gestures,' *Information and Computer Security*, 2022, ISSN: 2056497X. DOI: 10.1108/ICS-12-2021-0212.

[13] A. Darabseh and D. Pal, 'Performance analysis of keystroke dynamics using classification algorithms,' Institute of Electrical and Electronics Engineers Inc., Mar. 2020, pp. 124–130, ISBN: 9781728172835. DOI: 10.1109/ICICT50521.2020.00027.

[14] P. Bours and S. Mondal, 'Performance evaluation of continuous authentication systems,' *IET Biometrics*, vol. 4, pp. 220–226, 4 Dec. 2015, ISSN: 20474946. DOI: 10.1049/iet-bmt.2014.0070.

[15] B. Hassan, E. Izquierdo and T. Piatrik, 'Soft biometrics: A survey: Benchmark analysis, open challenges and recommendations,' *Multimedia Tools and Applications*, 2021, ISSN: 15737721. DOI: 10.1007/s11042-021-10622-8.

[16] T. A. Salthouse, 'Effects of age and skill in typing,' *Journal of Experimental Psychology: General*, vol. 113, pp. 345–371, 3 Sep. 1984, ISSN: 00963445. DOI: 10.1037/0096-3445.113.3.345.

[17] A. O. Adesina and O. Oyebola, 'An investigation on the impact of age group and gender on the authentication performance of keystroke dynamics,' *IJARCCE*, vol. 10, 9 Sep. 2021, ISSN: 23195940. DOI: 10.17148/ijarcce.2021.10903.

[18] J. H. Roh, S. H. Lee and S. Kim, 'Keystroke dynamics for authentication in smartphone,' Institute of Electrical and Electronics Engineers Inc., Nov. 2016, pp. 1155–1159, ISBN: 9781509013258. DOI: 10.1109/ICTC.2016.7763394.

[19] I. Tsimperidis, C. Yucel and V. Katos, 'Age and gender as cyber attribution features in keystroke dynamic-based user classification processes,' *Electronics (Switzerland)*, vol. 10, 7 Apr. 2021, ISSN: 20799292. DOI: 10.3390/electronics10070835.

[20] D. I. Belov and R. D. Armstrong, 'Distributions of the kullback-leibler divergence with applications,' *British Journal of Mathematical and Statistical Psychology*, vol. 64, pp. 291–309, 2 May 2011, ISSN: 00071102. DOI: 10.1348/000711010X522227.

[21] K. S. Killourhy, 'A scientific understanding of keystroke dynamics,' 2012.

[22] A. Cutler, D. R. Cutler and J. R. Stevens, *Random forests*, 2012. DOI: 10.1007/978-1-4419-9326-7_5. [Online]. Available: http://link.springer.com/10.1007/978-1-4419-9326-7_5.

[23] M. Awad and R. Khanna, *Support vector machines for classification*, 2015. DOI: 10.1007/978-1-4302-5990-9_3.

[24] S. Sharma, J. Agrawal and S. Sharma, 'Classification through machine learning technique: C4. 5 algorithm based on various entropies,' *International Journal of Computer Applications*, vol. 82, pp. 28–32, 16 Nov. 2013. DOI: 10.5120/14249-2444.

[25] S. Sperandei, 'Understanding logistic regression analysis,' *Biochemia Medica*, vol. 24, pp. 12–18, 1 2014, ISSN: 13300962. DOI: 10.11613/BM.2014.003.

[26] A. Mhenni, E. Cherrier, C. Rosenberger and N. E. B. Amara, 'Analysis of doddington zoo classification for user dependent template update: Application to keystroke dynamics recognition.' [Online]. Available: https://hal.science/hal-02050173.

[27] P. Bours and S. Mondal, *Continuous Authentication with Keystroke Dynamics*. Feb. 2015, pp. 41–58. DOI: 10.15579/gcsr.vol2.ch3.

[28] A. T. Kiyani, A. Lasebae, K. Ali, M. U. Rehman and B. Haq, 'Continuous user authentication featuring keystroke dynamics based on robust recurrent confidence model and ensemble learning approach,' *IEEE Access*, vol. 8, pp. 156 177–156 189, 2020, ISSN: 21693536. DOI: 10.1109/ACCESS.2020.3019467.

[29] Ananya and S. Singh, 'Keystroke dynamics for continuous authentication,' Institute of Electrical and Electronics Engineers Inc., Aug. 2018, pp. 205–208, ISBN: 9781538617182. DOI: 10.1109/CONFLUENCE.2018.8442703.

[30] H. Barghouthi, 'Keystroke dynamics how typing characteristics differ from one application to another,' 2009.

[31] M. J. Pedersen, 'Keystroke dynamics based text copying detection,' 2021.

[32] L. Aversano, M. L. Bernardi, M. Cimitile and R. Pecori, 'Continuous authentication using deep neural networks ensemble on keystroke dynamics,' *PeerJ Computer Science*, vol. 7, pp. 1–27, 2021, ISSN: 23765992. DOI: 10.7717/PEERJ-CS.525.

[33] L. Yang, C. Li, R. You, B. Tu and L. Li, 'Tkca: A timely keystroke-based continuous user authentication with short keystroke sequence in uncontrolled settings,' *Cybersecurity*, vol. 4, 1 Dec. 2021, ISSN: 25233246. DOI: 10.1186/s42400-021-00075-9.

Appendix A

Additional Material

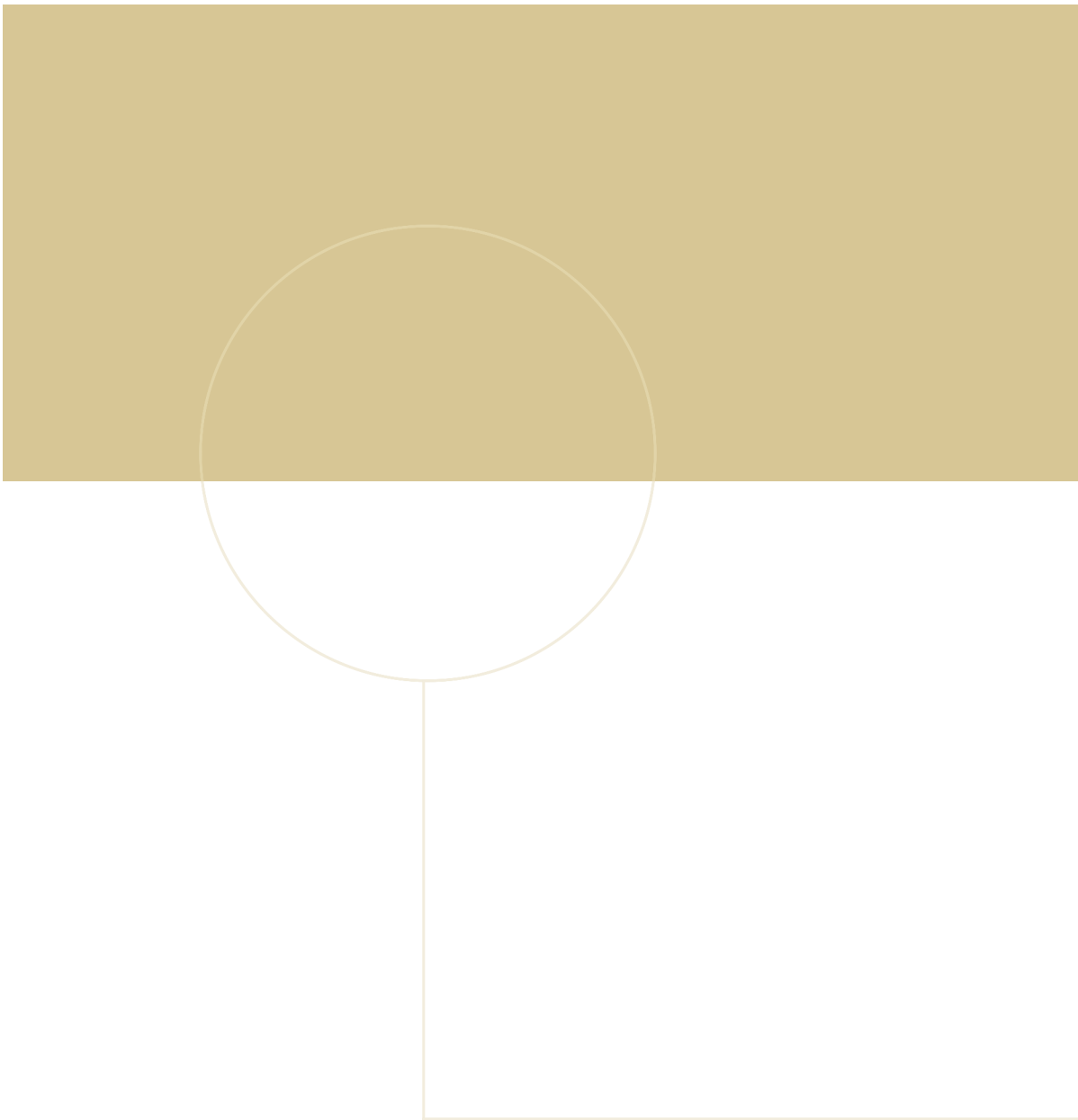
The following tables are the latency's and duration's for the manual feature selection tests. These are used in the tests where the feature selection method is not mentioned and consists of manually picked features by analyzing the pdf through bell curves.

RPlat
z -> i
c -> y
e -> h
CapsLock -> i
i -> u
0 -> Space
a -> h
b -> v
CapsLock -> Space
e -> r
k -> n
v -> Space
Space -> -
r -> h
CapsLock -> W
Space -> 1
B -> u
t -> p
M -> a
t -> :
Shift -> J
g -> t
t -> v
Shift -> K
Space -> 2
Space -> 3
b -> s
D -> o
Z -> i
CapsLock -> G
CapsLock -> a
W -> h
N -> Space
t -> n
H -> a
i -> n
H -> i
Shift -> ;
k -> ,
Shift -> F
e -> x
) -> Space
CapsLock -> T
CapsLock -> A
t -> Shift
x -> e
i -> l
O -> n
Shift -> (
h -> a
H -> e
a -> s
i -> s
r -> e
o -> t
a -> n
n -> d
i -> n
e -> Space
h -> e

Table A.1: Latency's

Duration
P
9
v
B
j
G
W
S
x
R
1
q
0
(
E
L
H
D
Control
C
U
Y
5
F
V
0
8
3
)
7
|
2
J
\
-
;
ArrowDown
e
.
n
w
c
M
I
K
t
?
A
:
k
h
CapsLock
,

Table A.2: Durations



NTNU

Norwegian University of
Science and Technology