Master's thesis

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

Cornelia Vedeld Plesner

Privacy vs. Security: Soft Biometrics in Distorted Keystroke Dynamics Data

Master's thesis in Communication Technology and Digital Security
Supervisor: Patrick Bours
June 2023

NTNU
Norwegian University of
Science and Technology

Cornelia Vedeld Plesner

Privacy vs. Security: Soft Biometrics in Distorted Keystroke Dynamics Data

Master's thesis in Communication Technology and Digital Security
Supervisor: Patrick Bours
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

Title: **Privacy vs. Security:** Soft Biometrics in Distorted Keystroke Dynamics Data

Student: Plesner, Cornelia Vedeld

Problem description:

Advances in technology and the increasing connectivity of devices and systems have transformed our way of living. Although this has undeniably brought innumerable positive benefits and opportunities, it facilitates new challenges concerning security and privacy. The attack surface increases and the need for security measures against intentional and non-intentional attacks grows. In addition, vast amounts of data are collected and stored every second, challenging the users' control over their personal information. Although privacy and security in many cases go hand-in-hand, there are areas of conflict, where enhancing one aspect might come at the expense of the other.

Not all individuals are familiar with the type and amount of data gathered, how it may be distributed, and how this might pose a privacy concern. Behavioral biometric Keystroke Dynamics is based on the idea that each individual has their own, unique way of typing on a keyboard. By collecting the timing information of the press and release of each key, it is possible to extract information about the individual, making it a suitable scheme for verification and both static and continuous authentication. Keystroke Dynamics provides another layer of security, due to the difficulty of impersonating how a user types.

It has been discovered that soft biometric features such as age and gender can be recognized in keystroke dynamics data with high approximate accuracy. Being able to uncover this may be beneficial in several areas, for instance, detecting deviations from targeted user groups of social media, e.g., in chat rooms for younger children and teens, where cyber grooming may be a risk. However, privacy concerns can be raised, as this data may be collected and a profile of the user can be built without their knowledge, for instance, for marketing or criminal purposes.

To protect the privacy of individuals' keystroke dynamics data, technical tools might be applicable to distort the data collected, such as the Google Chrome extension "KeyboardPrivacy" created by Paul Moore. This tool will interfere with the periodicity of the Keystroke Dynamics data collected, claiming it will be impossible to profile and/or identify the user.

With this in mind, the main research question of this thesis regards the feasibility to identify soft biometrics, such as age and gender, even when the collected keystroke dynamics data is distorted. The objective is to investigate how the noise influences the timing data and evaluate if it makes Keystroke Dynamics in combination with Soft Biometrics ineffective. The study intends to examine the trade-off between security and privacy so that it can provide a contribution to the ultimate aim of increasing safety online.

Approved on: 2023-02-15
Supervisor: Bours, Patrick, NTNU

Abstract

In today's rapidly evolving digital landscape, the preservation of privacy and security can be a daunting task. Utilizing keystroke dynamics to enhance authentication and identification techniques is a promising approach that increases security, but at the same time raises important privacy considerations to address. Hence, this thesis aims to investigate whether distortion of Keystroke Dynamics data can hinder the detection of soft biometric characteristics, such as age and gender.

A program was used to simulate and add distortion to the data in combination with the Google plug-in tool "Keyboard Privacy". The data underwent processing and subsequent analysis using the Machine Learning model Support Vector Machine in order to classify age and gender. Additional analysis was carried out to determine if it was possible to detect any distortions within the dataset.

The study revealed that there are distinguishable differences between distorted and non-distorted keystroke dynamics data. While the patterns may bear similarities, they are still distinct enough to enable relatively accurate classification. The performance of the distorted dataset may vary depending on the classification categories, where gender classification performed better than age classification. These findings shed a light on the possibility of developing more sophisticated systems for biometric identification and authentication.

Sammendrag

Det digitale landskapet utvikles raskt, noe som gjør bevaring av personvern og sikkerhet til en kompleks oppgave. Å utnytte Keystroke Dynamics for å forbedre autentisering og identifikasjon av individer har et stort potensial for å forbedre sikkerhet, men som samtidig bærer med seg bekymring rundt behandling av personvern. Denne oppgaven har som mål å undersøke om forvrengning av Keystroke Dynamics data kan forhindre deteksjon av Myke Biometriske kjennetegn, slik som alder og kjønn.

Et program, i kombinasjon med forvreningsverktøyet i Google; KKeyboard Privacy", ble brukt for å simulere og legge til forvrengninger på dataen. Deretter ble dataen kjørt inn i en maskinlæringsalgoritme, nærmere bestemt Support Vector Machine, for å klassifisere simulert og forvrengt data på alder og kjønn. I tillegg ble det analysert om det var mulig å detektere forvrengninger i datasettet.

Arbeidet har avdekket at det er merkbare forskjeller mellom de forvrengte og ikke-forvrengte datasettene. Selv om mønsterne bærer likheter, er de fremdeles ulike nok til å muliggjøre en relativt nøyaktig klassifisering. Ytelsen av de forvrengte datasettene kan variere på klassifiseringskategorien, der resultatene viser at klassifisering på kjønn har bedre ytelse enn på alder. Disse funnene belyser mulighetene for å utvikle mer sofistikerte systemer for biometrisk identifikasjon og autentisering.

Preface

This Master's thesis concludes my Master of Science in Communication Technology and Digitial Security within Information Security from the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway, in the spring of 2023. The study has been supervised by Patrick Bours and has been conducted by the Department of Information Security and Communication Technology.

First and foremost, I would like to express my sincerest gratitude towards my supervisor Patrick Bours. For being the most patient and understanding mentor, for great insights and invaluable guidance. For giving me a chance to dig deeper into such a fascinating subject that I never imagined I would come across. In addition, a special thanks to Tobias Moe, who has been very helpful, especially regarding the simulation phase. Thanks to family and friends, I am forever grateful for all of you.

Contents

List of Figures

List of Tables

List of Acronyms

D1 Distorted dataset 1.

D2 Distorted dataset 2.

IQR Interquartile Range.

KD Keystroke Dynamics.

KDA Keystroke Dynamics Authentication.

ML Machine Learning.

PW1 Password 1.

PW2 Password 2.

PW3 Password 3.

PW4 Password 4.

PW5 Password 5.

SVM Support Vector Machine.

Chapter 1
Introduction

Keystroke Dynamics has achieved promising results regarding the authentication and identification of users, using statistical and machine-learning approaches. This thesis explores the feasibility of classifying soft biometrics, specifically age and gender, even when the timing information of a user's keystrokes have been utilized.
The background and motivation behind this research are explained in the first section. Following, Section 1.2 present the objectives and research questions, including the hypothesis. Section 1.3 explains an important aspect to keep in mind when evaluating this study. A summary of contributions is presented in Section 1.4, and the structure of the thesis is given in Section 1.5

1.1 Background and Motivation

The rapid development of information and communication technology has had a profound impact on our society and made the intersection of privacy and security a crucial point of discussion. With the growing popularity of biometrics to enhance digital security, individuals are increasingly apprehensive about safeguarding their personal privacy.

The behavioral biometric, Keystroke Dynamics, captures an individual's unique typing pattern, where typing speed, duration, and latencies of keypresses are some of the characteristics that can be collected and analyzed. The technique has emerged as a promising alternative to traditional biometric solutions such as fingerprint scanning, particularly for those who seek a cost-effective and less intrusive approach. The only requirement is a standard computer keyboard; the keystrokes can be collected through software without the user noticing, not causing any disturbance to the user experience.

Several studies have achieved low error rates and high accuracy when identifying and authenticating users, implying a great potential for applications in the need of methods where efficiency is required. As the study of this proceeds, a more secure user verification can be established, while simultaneously becoming more user-friendly and adaptable.

Paradoxically, the capturing of keystroke dynamics of a user's typing behavior without the user's knowledge and consent can be perceived as uncomfortable and troubling to many. The concern grows increasingly evident, knowing that personal and sensitive information can be revealed and connected to the user's profile. With the enormous amount of data being collected, stored, and analyzed, the potential aftermath of data breaches increases, ranging from physiological distress to financial losses.

As a counteract to the privacy concerns, anonymization tools have been developed with the aim of protecting the users' data, such as keystroke dynamics data. KeyboardPrivacy is a browser plug-in that disrupts the typing data of the user, making it challenging to utilize keystroke dynamics to build a profile of the user. By adding random delays to the keystrokes, no pattern should be recognized, and the user's privacy will be preserved.

Although anonymization tools can enhance privacy protection, they also introduce new challenges. Users with malicious intents can misuse these tools to hide their keystroke dynamics data. For example, engaging in harmful activities online, such as committing financial crimes or cyber grooming, without the fear of being identified through their typing pattern. Focusing on the latter, different online arenas facilitate communication between individuals, exchanging knowledge, experience, and opinions. Some of these arenas have anonymous participation, where individuals have the opportunity to communicate freely without fear of repercussion. Unwanted and harmful circumstances can arise if certain sensitive information is shared with the wrong person. This exemplifies in chat forums for children, where predators can hide their identity, and thus exploit and manipulate minors for personal gains. By using an anonymization tool in addition to a fake identity, the predator conceals their digital footprint of keystroke dynamics, reducing the risk of detection.

The study's motivation is to identify user's soft biometrics traits, even when distortion tools are used, to ensure a safer online environment for children and aims to classify age and gender based on distorted keystroke dynamics data.

1.2 Objectives and methodology

The previously stated motives have brought forward the main research question, maintained from the project report as is [Ple22];

> **"Is it possible to identify soft biometrics, such as age and gender, even when the collected keystroke dynamics data is distorted?"**

To be able to substantiate the research further and gain a comprehensive understanding, the following sub-questions were defined in the project:

1. ***SQ1:*** *What difference does distorted and non-distorted timing data have when it comes to performance?*
   Due to different timing data, it is reasonable to imagine different results in performance. Are there any indications that prove the difference in values, and is there a pattern that can be discovered? It is necessary to know the differences in the statistical values and how the values are affected by a lower higher grade of distortion.

2. ***SQ2:*** *How should distorted data be handled?*
   When the data has been distorted, the techniques used for analysis must be customized accordingly. Best practice for non-distorted data may not work for distorted data, while the method for distorted data may be useless for non-distorted data. Is it possible to find a method that works adequately for both distorted and non-distorted data?

3. ***SQ3:*** *Is it feasible to detect whether the timing data collected is distorted? And if so, what kind of features differ?*
   It is necessary to determine if there is some kind of correlation between the distorted datasets. What features of the timing data differ, and how can this relate to the original data so that better analysis can take place?

Hypothesis:

The hypothesis remains consistent with the one proposed in the pre-project; *Soft biometrics characteristics that can be obtained from Keystroke Dynamics is, to a certain extent, distinctive. Thus, some level of distortion of the timing information will not be preventative for identification.*

1.3 Disclaimer

The research presented on Keystroke dynamics and the potential of classification on gender in this thesis has employed a binary gender model; male or female. This is due to the available data on these categories. The study acknowledges and respects the gender identities beyond this and has no intention to invalidate individuals or cause any harm.

1.4 Contributions

To summarize the findings in this thesis, the contributions with the most significance are as follows:

- The finding that the use of Interquartile Range (IQR) can potentially serve as a method for detecting distorted datasets.
- The finding of using mean values of latencies and durations as features when using distorted datasets in classification. This is due to the averaging of values tending to out-level noise, leading to improved performance.
- The finding of the potential value in exploring unsupervised machine learning approaches to the classification of soft biometrics to match real-world scenarios.

1.5 Structure of the thesis

This Master's Thesis consists of 8 chapters, and is organized as follows:

- **Chapter 2:** Provides the reader with the necessary background theory surrounding the relevant topics.
- **Chapter 3:** Provides a State-of-the-Art on user profiling and classification in Keystroke Dynamics.
- **Chapter 4:** Presents the GREY-NISLAB Keystroke Dynamics dataset and the processing and simulation of this.
- **Chapter 5:** Provides the methodology of the analysis performed and explanations of the metrics.
- **Chapter 6:** Presents the results retrieved from the analysis.
- **Chapter 7:** Evaluates the achieved results and discusses the findings with the research questions in mind.

– **Chapter 8:** Conclusion of the thesis with final remarks on the conducted research, contribution, and further work.

Chapter 2
Background and State of the Art

2.1 Information Security and the CIA Triad

The CIA triad is commonly considered the center of the entire information security discipline. Since the beginning, it has been the industry standard to help organizations identify and prioritize their security needs and measures [Sta15; Osc03]. It is constructed out of the three fundamental objectives in information security; **C**onfidentiality, **I**ntegrity, and **A**vailability.

- **Confidentiality:** ensuring information protection so that no information is exposed or disclosed to unauthorized individuals.
- **Integrity:** ensuring information will not be exposed for unauthorized modification.
- **Availability:** ensuring information is available and accessible for authorized individuals when necessary.

The field is continuously evolving, and there is a general consensus that these three terms fall short of fully encapsulating the security concepts. Stallings introduces the additional concepts of Authenticity and Accountability for a more holistic view.

- **Authenticity:** ensuring the information comes from a trusted and authorized source. It is closely connected to the objective of integrity but differs by focusing on the verification of the origin and not the state of the information.
- **Accountability:** ensuring an individual, organization, or system takes responsibility for their actions regarding the information they have access to.

This thesis primarily surrounds ensuring user authenticity, typically acquired through identification and authentification.

2.1.1 Privacy

Our lives are more intertwined with technology than ever, and vast amounts of data and information are stored in databases worldwide. Some of this data is considered personal data, and some might be classified as sensitive. With multiple platforms and stakeholders having the opportunity to collect and store this data, it is vulnerable to misuse.

According to the definition of personal data by the EU, any information that pertains to an individual who can be identified, either directly or indirectly, is considered as personal data. Examples of this include, but are not limited to, name, social security number, financial information, health information, and biometric data [GDPR16]. The tolerance and boundaries of required security levels are determined by the type of information. Some personal data is regarded as sensitive information, which may cause harm if disclosed. Therefore, handling sensitive information requires a higher level of security, ensuring that the objectives of privacy and confidentiality are accomplished.

Privacy is defined as the ability of an individual to manage the gathering, storing, and utilization of their personal data, as well as the parties with whom it is shared [Sta15]. A wide range of threats against personal data is present, such as identity theft, stalking, or surveillance, which makes it essential to protect data. Furthermore, companies can collect data for analysis and utilize this for their own benefit, such as improving their products and services, optimizing processes, and creating targeted and personalized marketing to increase sales and revenue. Some users may view the utilization of their data as intrusive and manipulative, essentially violating their privacy. As a result, the awareness and desire to safeguard their data have increased over the last few years, resulting in law enforcement and policies for data collection, such as the General Data Protection Regulation (GDPR) in the European Union and the California Consumer Privacy Act (CCPA) in the United States.

2.1.2 Data distortion through Keyboard Privacy

Keyboard Privacy is an extension in the Google Chrome browser with the goal of ensuring the user's privacy of Keystroke Dynamics data [Moo15]. It was created as a counteract against invasions of privacy and seeks to prevent the profiling of individuals based on the typing information provided when using the keyboard. "Keystroke Dynamics will not be useable for authentication purposes, as the input data won't match the template data.", Moore states. Testing the plug-in on *KeyTrac*, a product that aims for biometric profiling of knowledge-based authentication, the original recognition rate of 82 % plummeted to 3%.

When the plug-in is activated, it will intercept the keyboard entry before it reaches the Document-Object-Model (DOM), adding a randomized delay to the durations and latencies. The delay added can be customized by the user for their preferences, but the default settings introduce a random delay of values between 0 and 200 milliseconds for both the duration and the latencies. A further look into the code reveals that only 50% of the features are added a delay [Moe21b].

2.1.3 Authentication

Authentication refers to the process of verifying that the communication between entities is authentic and legitimate [SRF19]. Fulfilling this objective ensures that only authorized parties are given access to information or systems. There are various ways for authentication, but they can generally be divided into three categories;

- **Knowledge-based:** based on the knowledge that individuals possess, such as passwords.
- **Possession-based:** based on the items an individual possesses, for example, an access card.
- **Biometric-based:** based on physiological attributes and behavioral characteristics of an individual, for example, a fingerprint or voice.

Password-based authentication, a scheme based on knowledge, remains the most common method, yet has frequently been compromised due to easy-to-guess passwords, re-use for multiple services, and poor password management, among others [For20]. As the technology continues to evolve, basic attacks such as brute force have been efficient for password cracking. It is not likely that passwords will not disappear any time soon due to their widespread use. However, the need for stronger authentication schemes has been taken seriously, adding layers of security by combining different methods, such as Multi-Factor Authentication and Single Sign-On.

Authentication can either be performed statically or continuously [GER11]. Static Authentication means that the authentication is only done once at the beginning of the session. The user will remain authenticated for the whole duration of the session and will not need to prove authenticity later on. Continuous authentication denotes the approach of authentication where the system will regularly need re-authentication throughout the session. Passwords or tokens do not come with the possibility of re-authentication; biometrics is needed in this case.

2.2 Biometrics

Biometrics is a term referring to the process of quantifying and analyzing individuals' unique characteristic traits and attributes that can later be used for identification, verification, or authentication [Nat17; JR08]. These can be categorized into two systems, *physiological biometrics* and *behavioral biometrics*. Physiological biometrics involves the physical composition of a human being, i.e., DNA analysis, fingerprint recognition, or facial geometry recognition. Behavioral biometrics involves the behavioral patterns of a human, essentially focusing on the distinct ways users perform and interact with their bodies. Examples include voice recognition, walking gait recognition, and typing pattern recognition, known as *Keystroke Dynamics*. Furthermore, the two categories differ in evolution over time, where the behavioral traits can, to a higher degree, be influenced and trained by the surrounding environment and external factors, whereas physicological traits evolve at a relatively slow rate over time.

A certain set of properties should be satisfied when utilizing biometric characteristics to differentiate users, listed in descending order of priority [DG04; Dor18].

- **Universality:** The biometric character should be widely available, and its application should be universal, without significant variations or limitations.
- **Distinctiveness:** The biometric character should be easily differentiated between two users.
- **Permanence:** The biometric character should be persistent and consistent over time.
- **Performance:** The biometric character should be recognized rapidly and accurately.
- **Collectability:** The biometric character should easily be collected and measured.
- **Acceptability:** The biometric character should have a high degree of acceptance amongst users, not causing any distress or discomfort.
- **Circumvention:** The biometric character should be difficult to mimic or spoof.

Ultimately, the most highly desired characteristics are universality, distinctiveness, and permanence. Universality is crucial to ensure that the usage won't be limited to a smaller set of individuals, distinctiveness is critical to provide accurate identification, and permanence ensures a reliable and accurate system over time. Utilizing biometric characteristics that lack a high level of these properties may reduce or compromise the accuracy and reliability of the biometric system, resulting in an increased number of false positives and negatives.

Biometrics offers several advantages, such as a seamless and efficient authentication process, where the user has to provide minimum effort and time to be accepted, making the technology ideal for situations where time is highly regarded and efficiency is required, such as financial transactions. Behavioral biometrics can be considered less intrusive, as it involves observing and analyzing the behavior over an extended period, which may be achieved without their awareness. Physical biometrics provide a higher level of accuracy as the traits show more stability over time, and their uniqueness will provide a more consistent basis for detecting impostors. Ultimately, biometrics are difficult to replicate and provide a secure and efficient way of identification.

Despite the advantages of accuracy and efficiency, the implementation of biometrics raises concerns regarding data handling, security, and privacy. Several biometrics can be considered sensitive information, facing risks of misuse and theft. Stored data is vulnerable to disclosure to unauthorized persons and requires proper management. In addition to the requirement of security measures, the systems are expensive to develop and maintain. A key point to consider is the risk of falsely authorizing and identifying individuals, false positives and negatives may still occur. Ultimately, the collection and handling of personal data raises ethical questions and concerns.

2.2.1 Soft Biometrics

The attributes contributing to the shaping of the behavioral and physical characteristics in individuals, such as age and gender, are known as *Soft Biometrics* [GER11]. These attributes cannot fully distinguish one person from another, resulting in it being a low discrimination attribute. This implies that the information associated can remain publicly accessible without compromising privacy. The characteristics will aid the identification and authentication process, enhancing accuracy and efficiency. Another advantage is that information collection is less invasive and requires less direct contact and equipment. However, this poses a privacy concern, where unauthorized parties have the ability to collect information without the acknowledgement or consent of the owner.

2.2.2 Multimodal Biometrics

The combination of two or more biometrics techniques, referred to as multimodal biometrics, can significantly improve the performance of recognition and authentication of biometric systems. By combining the use of a fingerprint scan with a face recognition system, the accuracy of correctly identified individuals can increase. Furthermore, these biometrics are at risk of presentation attacks; however, the complexity of an attack immediately increases when multimodal techniques are taken into use, hence an increase of security [AYRA21].

A study from 2017 presented a biometric system predicting age and gender, combining Keystroke Dynamics with Mouse Dynamics. They achieved better predictions than chance for all the created models, where some of the supervised machine learning models yielded f-scores up to approx. 0.9, predicting both the age and gender of the user[Pen17].

There are various methods of how multimodal biometrics can be implemented. Relevant for this thesis is *score fusion* and *decision fusion*. Other methods will not be mentioned.

- **Score fusion:** Individual scores from each modality are calculated and later combined into one single score to determine one final score. There are several methods, such as weighted average and maximum likelihood estimation [RN09].
- **Decision fusion:** Decisions taken by each modality are combined into one single decision by following predefined rules or algorithms [RMT21].

2.3 Keystroke Dynamics

KD is a behavioral biometric that measures and analyzes the typing behavior a user exhibits on a keyboard, making it possible to build a profile of the individual[MR00; GER11]. Raw timing information about the press and release of keys are gathered and later utilized to extract different features, such as latency and duration, which are the most commonly used [TA20]. Through the implementation of statistical models and machine learning algorithms, identification of the distinctive typing patterns of individual users is possible. The algorithms will be able to detect unique characteristics so that a biometric template can be used for various applications, such as user authentication and verification.

KD research and usage are gradually growing due to its inexpensiveness, low computational complexity, and ease of use [RPK+22; Sye14; Bou12]. It does not require additional hardware beyond a keyboard, and the user does not have to perform any specific action besides typing on the keyboard. There are diverse potential application areas for KD, such as data security and access control, where KD authentication has provided promising results, with accuracies up to 98% [ACB18]. It can contribute to evidence collection and profiling of cybercriminals; it has been shown that age, gender, and handedness can be predicted with a conventional keyboard [Sye14]. KD has also been used for the recognition of a user's emotional state and has been able to detect emotions such as joy and fear, relevant for the health sector [NAMH14; Epp10].

Several challenges with KD-based systems are present. Difficulty with data acquisition, user variability, and external factors makes KD more unreliable. To obtain accurate results, data collected must be of high quality, which depends on hardware and software specifications, method of collection, and environmental factors such as surrounding noise. Various factors, including physiological and cognitive conditions, familiarity with the keyboard layout, and personal qualities, can influence a user's typing pattern. Typing on the keyboard is not a static trait but rather a skill that improves precision over time, which can be challenging to capture [RPK+22; LAS17; Epp10]. These conditions may impact the typing behavior, causing variations in timing and pressure, leading to reduced quality of results.

2.3.1 Data capturing

A single session can generate thousands of keystrokes, allowing for extracting features that can subsequently be utilized for profiling [GER11; SCRB14]. Other features can be calculated and extracted by the raw data, including value, event, and timestamp, further explained in Section 2.3.2.

- *Value:* Each key on the keyboard has its own value, also denoted keycode. It incorporates the key's location on the keyboard and provides support for differing keys that have the ability to produce the same characters. It is commonly represented as an ASCII code, but other formats are possible, depending on the keyboard, operational system, and language.
- *Event:* the type of action that has occurred on the keyboard, either *press* or *release*. The timing of a key pressed down is noted as Press Time, or (Key) Down Time. The timing of a key being released is noted as Release Time, or (Key) Up Time
- *Timestamp:* the timing of the occurring event, normally counted with milliseconds, but the unit is not enforced.

2.3.2 Feature selection

The features of most significance are the duration and the latencies of the events [RPK+22]. Other features like force, pressure, and sound might be applied. However, this is outside of the scope of the thesis, and will not be examined further. The features are illustrated in Figure 2.1.

- *Duration*: The period a key is being held down, also noted as hold time or dwell time.
- *Latency:* The time delay between the press of a key and the subsequent key being pressed, also noted as flight time. Different types of latencies can be calculated, which can be positive and negative, as one key can be pressed before the previous key is released.
  - **Press-Press-Latency:** Latency between the press of one key and the subsequent key.
    $$lat_{pp} = dur_{currentKey} + lat_{rp}$$
  - **Release-Release-Latency:** Latency between the release of one key and the subsequent key.
    $$lat_{rr} = lat_{rp} + dur_{nextKey}$$
  - **Press-Release-Latency:** Latency between the press of one key and the release of the subsequent key.
    $$lat_{pr} = dur_{currentKey} + lat_{rp} + dur_{nextKey}$$
  - **Release-Press-Latency:** Latency between the release of one key and the press of the subsequent key.

2.3.3 Fixed/Free Text

KD can be applied to either fixed text or free text [SCRB14].

- *Fixed text:* Refers to a predefined text that the user will be required to type several times to be able to be authenticated or identified, such as a password or passphrase. This text will not change. Hence the system can be trained to recognize this specific manner of typing for the specific text. This is done using statistical and machine learning approaches to build a profile of this behavior.
- *Free text:* Refers to longer portions of text that are not predefined - the user is able to type anything they desire at any given time. The system extracts features to build a profile of the user and will later use this to compare the extracted features from the same or another different text.

Both of these approaches have advantages and disadvantages. The fixed-text approach requires less storage and resources, and users may have more consistency in their typing. This facilitates simpler analysis and more accurate results. However,

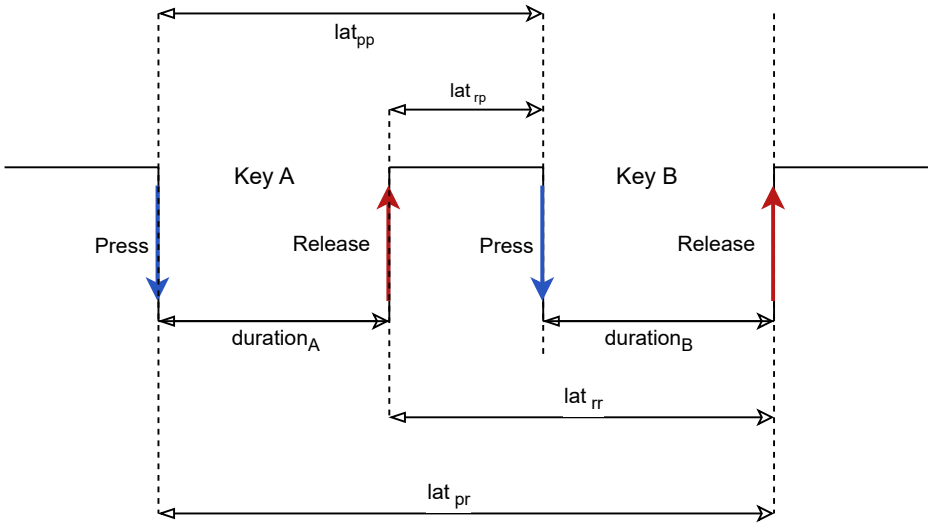


Figure 2.1: Illustration of the features in KD. Illustration derived from Morales, Aythami and Falanga, Mario and Fierrez, Julian and Sansone, Carlo and Ortega-Garcia, Javier[MFF+15].

it can be found less convenient by the user and may be more vulnerable to attacks like imitation and forgery.

On the other hand, the free-text approach provides a more natural environment for the user, and replicating the typing behavior is more challenging. This makes it a good scheme for situations where a change of behavior must be acted upon, such as removing access. However, the continuous monitoring of the behavior might feel like an invasion of the user's privacy, and the scheme requires more collected data and more complex analysis.

2.4 Reference Template

For each user to be recognized by its typing behavior, a reference template is created. This is created by utilizing the extracted features from the KD data that has been collected during the user's interaction with the keyboard. To prevent the occurrence of false admissions or rejections throughout the authentication process, it is essential that the template accurately represents the typing behavior of the user and is updated over time.

The reference template creation is divided into an enrolment phase and an authentication phase [Bou12]. The process differs slightly depending on the authentication type, which can be static or continuous. The authentication types are elaborated

further in section 3. When initially typing a phrase, such as a password, there will potentially be significant variations in typing rhythm. This might be due to the unfamiliarity of typing a new password or other external factors. During the enrolment phase, the data collected will serve as a reference template until the user becomes comfortable with typing their password. As a result, initial keystrokes will not be captured until the typing pattern becomes more stable. Once the typing is consistent, the data is collected and utilized for the construction of the reference template. This is somewhat different from other biometrics, such as fingerprints or iris, where the enrolment phase is immediately established [RPK+22]. More details about template creation will be provided in the sections 3.2 and 3.1.

2.4.1 Keystroke Dynamics and Soft Biometrics

As mentioned, KD is a behavior-based biometric that uses the individuals' unique way of typing to identify or authenticate the user. Soft Biometrics are non-intrusive biometric traits, easily acquired, and contribute to the same objective. Therefore, combining these fields can provide essential insights into behavioral biometrics and the factors influencing the variations in typing behavior. The field is relatively unexplored, however, there are studies with promising results, providing the potential to yield more comprehensive and secure biometric methods.

Research suggests that the age group and gender of an individual can be predicted, based on their vocabulary, stylistics, and personality [Bon21]. Furthermore, it has been found that the physical patterns differ between the age groups, revealing new valuable information, and making it possible to predict the age of a user. A study from 2017, exploring the prediction of age based on KD in combination with mouse patterns obtained an f-score of 0.62, with a best result of 0.92 in the age group 16-19 [Pen19]. The f1-score is the harmonic mean between the precision and recall, further explained in Section 5.2.4.

Gender prediction has also been showing promising results. An infographic by *Ratatype* claims that boys generally type faster than girls, whereas boys write with the typing speed of 44 *words per minte (wpm)*, while girls type with a speed of $37wpm$ [Rat]. A study from 2019 found that the gender of a random user on the Internet can be identified with an accuracy of 95.6% with a few hundred features [TA20]. They found that the average value of all digram latencies of males was $373.04ms$, and the women's average value of all digram latencies was $375.71ms$, However, with a standard deviation of $135.26ms$ for males, and a standard deviation of $116.86ms$ for women, it was concluded that women tend to have a more consistent typing pattern when analyzing the digram latencies, and with machine learning, the system was able to predict the correct gender 78% of the time. The same study examined age classification, however, somewhat lower results were achieved with an

accuracy of 74.2%, when using a Support Vector Machine model with a Polykernel and C parameter of 0.5. Using a Radial Basis Function Network, the accuracy was increased to 89.2%. Another study performed by Roy, Roy, and Sinha in 2018 achieved accuracies of 83%-95% for gender classification using Fuzzy-Rough Nearest Neighbor with Vaguely Quantified Rough Set (FRNN-VQRS), a machine learning technique utilizing distance metrics. The same study achieved accuracies on age classification of the range 75% to 94% [RRS18].

In a study performed by Fairhurst in 2012, a gender prediction accuracy of 80 % was achieved, and Syed achieved an accuracy of 91.6% for users under the age of 15, 2 years later[FD12; SCRB14]. Syed included prediction of handedness, whether a user is left- or right-handed, and achieved an accuracy within the range of 85% to 92% by using fixed text KD and fusion.

2.5 Machine Learning

Machine Learning (ML) is a subset within data science where systems and models are trained to perform specific tasks, for example classification or clustering. Rather than being explicitly told how to carry out these tasks, the systems identifies patterns and correlations by 'learning' from the provided training data.

These tasks, which may include classifications, are performed by these systems using learning algorithms and training data sets. Rather than being explicitly programmed to carry out these tasks, these systems 'learn' to do so by identifying patterns and making inferences from the provided training data [YLH03; Mag07; Qin20]. As a result, the machines can more effectively and efficiently interpret large amounts of data, extract relevant features and information and make decisions based on this. In the later years, ML has been actively used for KD systems. By learning the unique typing patterns of individuals, such as latency and duration of keypresses, the information can be utilized for different purposes, such as user authentication and verification. ML is commonly divided into three paradigms, unsupervised learning, supervised learning, and reinforcement learning. The latter will not be discussed further.

2.5.1 Unsupervised Learning

Unsupervised learning regards the training of a system where labels are not provided to the features. This way, the machine is forced to find patterns, correlations, and structures in the data without any kind of guidance. Typical applications are clustering and association of data.

2.5.2 Supervised Learning

Supervised learning regards the environments where the system is provided with features and their corresponding labels during the training phase. The system will then be able to predict labels when provided with unforeseen data. Known applications are classification and regression.

2.5.3 Support Vector Machine (SVM)

The supervised learning algorithm Support Vector Machine (SVM) is used for classification and regression tasks, and has a wide range of deployment, such as text categorization, bioinformatics, face detection, and more [YLH03]. Not relying on distance metrics or assuming a specific distribution of the data, including a certain robustness to outliers, makes the SVM a flexible solution.

The main objective of an SVM is the pattern recognition and separation of two or more classes during the training phase by determining a surface that divides them [CGRL20; Kec05]. This surface can be referred to as a hyperplane due to the possibility of the feature space being in infinite dimensions. The SVM aims to calculate a surface that maximizes the minimum distance between the classes while all the data points stay on the right side of the hyperplane relative to their group. Hence, the nearest data points are referred to as *Support Vectors*. A kernel function calculates similarities between different data point pairs and projects these relationships to a higher-dimensional space, separating these with a hyperplane, thus classifying the data. Figure 2.2 illustrates the separation of classes by a hyperplane in 2-dimensional space, with the support vectors creating the maximum distance.

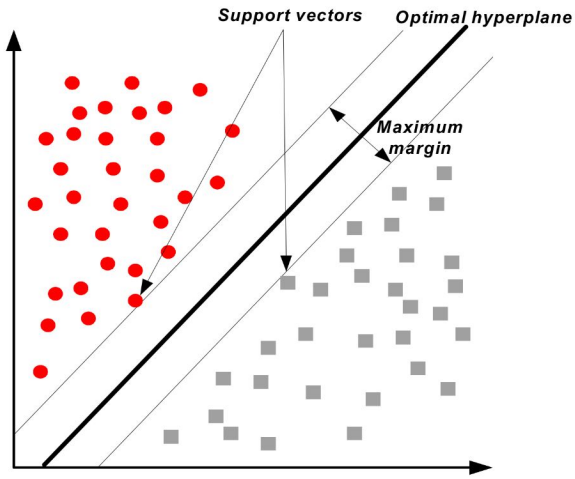


Figure 2.2: Illustration of SVM, fetched from [CKH+19]

KD information is usually gathered as high-dimensional data containing a large number of features, which can be challenging to separate linearly. SVM efficiently handles this by mapping the data to a high-dimensional feature space, making the data more separable.

2.5.4 Hyperparameter tuning

To achieve the optimal performance, the choice of hyperparameters is crucial. These are configurations that are not learned from the training process but rather act as guidelines for how the model should process and learn the information. There are several different ways of optimizing the choice of hyperparameters, such as grid search, random search, and Bayesian optimization, all with the purpose of selecting the combination where the best results are achieved. Grid Search is a method where all possible combinations of the hyperparameters are used to train the model, cross-validated, and later compared against each other based on accuracy.

Choice of kernel

The selection of the kernel function has a significant impact on the performance, as different data require different alterations. The choice depends on the type of data and how this should be transformed. Two common kernels are:

- **Radial Basis Function Kernel (RBF):** The similarity between two data points is calculated as a function of their original distance in the transformed feature space. This decreases exponentially.
- **Polynomial Kernel:** Examines the similarity between the data points, including different combinations of these [Wik22].

Cost Parameter

The cost parameter is relevant to achieve the lowest error rate while keeping the model complexity to a minimum. The cost parameter assigns penalties for wrong predictions, where a high cost strongly encourages the model to accurately classify the training data by assigning heavy penalties if there is misclassification. This enhances the performance, however, the complexity of unseen data rise and carries a risk of overfitting. A lower cost, on the other hand, allows a larger margin for the toleration of misclassification. This may lead to underfitting, as the underlying patterns in the KD data won't be captured.

Gamma parameter

The shape of the decision boundary is controlled by the Gamma parameter and how much each point influences the drawing of the hyperplane. A low value of gamma

results in a linear, less complex model, with a broader decision boundary, where the data points further away is considered. Conversely, a high gamma results in a more complex, non-linear decision boundary, where the data points closer to the boundary line have a more significant influence and can capture a more profound, underlying pattern. However, a higher value increases the risk of overfitting, collecting noise in addition to the underlying pattern.

2.5.5 Overfitting

When the model learns the data from the training set too well, it begins capturing noise and random fluctuations [Kec05; YLH03], referred to as *overfitting*. Figure 2.3 illustrates two instances of a simple classification task, where the two classes are represented by squares and circles. The left figure shows a perfect classification, where the training data are separated into their correct class. The right figure has different hyperparameters, leading to a different separation boundary, causing the model to overfit. The filled circles and squares are the wrongly classified samples, showing that the model misinterpret the true pattern of the data and fails to generalize the data. *Underfitting* is the case where the machine cannot learn the underlying pattern in the training data, and hence fails to generalize new test data.

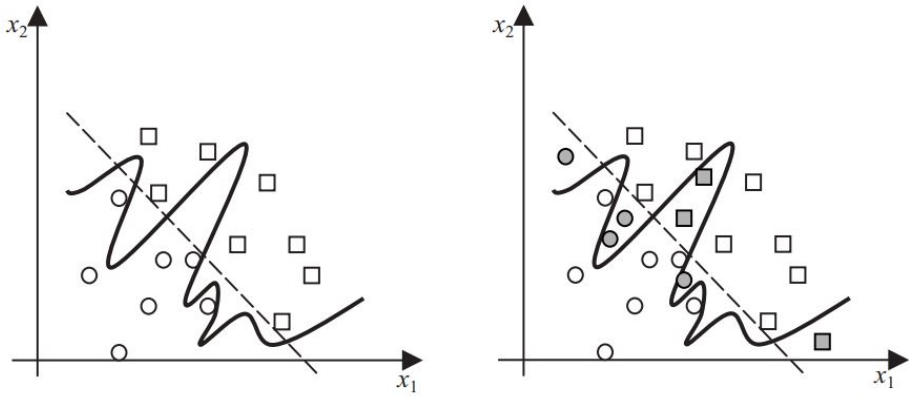


Figure 2.3: *Left:* The case of perfect classification. *Right:* The case of overfitting [Kec05]

Chapter 3
User Profiling and Classification

This chapter explains the concepts behind the profiling of users for authentication and classification using KD.

3.1 Keystroke Dynamics for Static Authentication

KD for Static Authentication is the simplest form and compares the keystroke features of a user once at the beginning of the session [Bou12]. This can be used for access control as an entry point, for instance, when the user types their username and password. This scheme considers a fixed text, and the typing is usually determined with distance metrics, such as Scaled Manhattan Distance. The distance value, d, will be compared to the threshold value, T. If the distance value is smaller than the threshold value, $d < T$, the user will be accepted, and access will be granted.

A distance measure estimates the distance between two data points [Sha20]. In this context, an entry will be validated against a model, a vector will be built, and this vector will be used to calculate the distance between the entry and the stored point.

Some distance measures have shown to be more effective than others in Keystroke Dynamics Authentication (KDA) [Dor18; Bou12]:

- **Manhattan Distance:** $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$
- **Scaled Manhattan Distance:** $d(x, y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{s_i}$
- **Euclidean Distance:** $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$
- **Canberra Distance:** $d(x, y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$
- **Chebyshev Distance:** $d(x, y) = \max_i |x_i - y_i|$
- **Mahalanobis Distance:** $d(x, y) = \sqrt{(x - y)^T S^{-1} (x - y)}$

where,

- x and y are two data points in the n-dimensional space, represented as a vector
- s_i is the scaling factor for the i^{th} dimension.
- S: the covariance matrix containing the interdependencies between the different dimensions.

In the authentication phase, a user will try to log in by typing their password once. From this typing, the relevant features, often durations and latencies, will be extracted and are represented by t_1, t_2, t_{2N-1}. Furthermore, these features will be compared to the reference template, and based on the predetermined criteria, the user is either rejected or accepted. These decision-making criteria are based on different thresholds between the distance from reference template T and the new input, probe I.

$$d(T, I) = \sum \frac{|t_i - \mu_i|}{\sigma_i} \tag{3.1}$$

A limitation of the static scheme is that there is no confirmation of true identity after initial authentication. After the initial user is authenticated, no re-confirmation of the user's identity is conducted. If there is a change of user later in time, this will not be detected, meaning that an unauthorized user can be able to gain access to systems or perform actions. Furthermore, static authentication is more susceptible to brute-force and impersonation attacks, as an attacker will have the opportunity to attempt to impersonate the typing rhythm repeatedly.

3.1.1 Static authentication templates

The reference template can be calculated by averaging the different values of durations and latencies, either by, for example, finding the mean or the median. Assume we have a password of N keys, meaning there are N durations and $N-1$ latencies. The reference template will consist of $2N-1$ pairs of (μ_i, σ_i). When the μ_i value is small, the user is typing with high stability, while a larger value represents less stability [Bou12]. The mean is considered a better option when regarding central tendency, while the median is better at handling outliers [Bou12].

3.2 Keystroke Dynamics for Continuous Authentication

It is possible to use KD to continuously monitor the user's behavior, making it possible to perform actions if the behavior has noticeably changed. For continuous authentication, the behavior of the user is monitored over a period of time or

throughout a session, where the user writes a free text [Bou12]. The comparison will no longer be with the fixed text but with different combinations of letters, *n-graphs*. This way, the identity will be re-confirmed after pre-determined time periods, either periodically (*Periodic Authentication*) or after every single keystroke (*Continuous Authentication*). However, this requires more storage and resources than static authentication and is more difficult to process than static KDA.

In continuous authentication KD, the timing of detecting an impostor is more important than if the impostor is detected at all [Bou12]. It is essential that genuine users will be kept from being locked out, and simultaneously lock out impostors as fast as possible, in other words, with the least amount of keystrokes possible. The later detected, the more damage can be done.

3.2.1 Continuous authentication templates

Continuous authentication is more complex, as it is unknown what the user will type. Therefore, the enrolment phase will need a time-based period to monitor typing behavior, e.g., for 1 or 2 days. With the help of di- and tri-graphs and their characteristics, a reference template can be created [Bou12]. N-graphs are different combinations of N letters, i.e., di- and tri-graphs, where di-graphs contain two letters, such as "*ke*", while tri-graphs contain three letters, such as "*key*".

The choice of letter combinations is of interest. Not all combinations must be included in the reference template; the number of times a letter or n-graph is typed during the enrolment phase needs to be high enough to be able to calculate statistical values, such as mean and standard deviation.

The continuous authentication scheme evaluates if a user is genuine, and the decision is re-evaluated after every keystroke. A user cannot be thrown out based on one single keystroke; however, if their typing pattern is significantly changed, action must be taken. To examine this, the *concept of trust* is introduced. After a successful static authentication procedure, the level of trust, C, will be determined, where the minimum value 0 represents no trust and the maximum value 100 represents complete trust. With the use of a *penalty-and-reward* function, the levels of trust will be affected. Similarly, as in the static KDA, the distance metrics will be calculated when the user types a key or a key combination.

Assuming the user types one key, k, the mean, $\mu_{dur,k}$, and the standard deviation, $\sigma_{dur,k}$, of the duration of the key press will be calculated and used in the reference table. Given the duration, $t_{dur,k}$, from the reference table, the distance between the probe and the reference table, D, can be calculated as:

$$D = dur_{dur,k} = \frac{|\mu_{dur,k} - t_{dur,k}|}{\sigma_{dur,k}}$$

With the calculated distance, D, the user will either receive a penalty or reward, dependent on the distance from the determined threshold, $T_{distance}$. This can either be an increase or decrease in the value C, with a constant, such as 1. Other variations might be to reward the user with a higher value than the penalty, or to introduce a variability based on the threshold and the distance, $\Delta = (T_{distance} - D)$. Δ will be either added or subtracted from the trust level, meaning that the higher the accuracy of the typing, the better reward, and vice versa. Ultimately, if the user obtains a C-value beneath the determined threshold, the user will be locked out.

3.3 Keystroke Dynamics Classification

Prediction and classification in KD utilize the knowledge-discovery model to predict the belonging of samples to classes [RPK+22]. The knowledge-discovery model is a model aiming to extract high-level knowledge from low-level data, by cleaning, processing and analyzing the data [CSPK07]. They are dependent on the samples of multiple users to be able to discover underlying patterns for the different classes.

To discover these underlying patterns and characteristics, machine learning is a commonly used method. Some of these include [RPK+22; NFG+22]:

- **Support Vector Machines:** A boundary between the classes is made, with the aim of maximizing the distance between the different classes.
- **Tree-based:** Include models like Decision Trees and Random Forests. These models obtain a series of tests or questions and make decisions based on these in a tree-like path.
- **Neural Networks:** These models recognize and learn patterns by interpreting sensory data.
- **Fuzzy Logic:** Fuzzy meaning 'not clear', this method is making decisions based on situations that are ranked based on their 'level of truthness', or level of uncertainty.
- **K-Nearest Neighbour:** The classification is done based on the similarity of the known data points, usually measured by distance functions.

This thesis utilizes Support Vector Machine, which is a common choice in KD due to its solid mathematical foundation and efficient classification. It is further described in Chapter 2.5.

Chapter 4
Data and data processing

This chapter will provide an overview of the collection of the data used for analysis, processing, and justification for the choices made. A benchmark dataset from the University of Caen Basse-Normandie and NISlab Gjøvik have been used. Later, the processing of the dataset through simulation has been presented. The simulation is not in the scope of the thesis, thus, only a brief explanation of the approach is included.

4.1 Data Collection

The research of KD data is known to be challenging due to the difficulty of obtaining enough data to give statistically significant results and providing high-quality data sets. The process of creating high-quality data sets can be time-consuming and challenging, as well as privacy and ethical concerns must be considered. The collected datasets will be run through a simulator to produce new ones with new characteristics. In addition to the original one, there will be three simulated datasets of each dataset to be analyzed, where two have been employing the Keyboard Privacy plugin. These have recieved randomized delay of $0-200ms$ and $0-300ms$ distortion in the duration and latency timing.

- **Original dataset:** referring to the unmodified dataset directly from the source.
- **Simulated dataset:** referring to the resulting dataset of the simulation of the original dataset.
- **Distorted dataset:** referring to the result of the simulation of the original dataset with enabled Keyboard Privacy plugin.
    - **Distorted 1:** Will also be referred to as Distorted dataset 1 (D1). Added a random delay of $0-200ms$ in duration and latency.
    - **Distorted 2:** Will also be referred to as Distorted dataset 2 (D2). Added a random delay of $0-300ms$ in duration and latency.

Analyzing both the original and the simulated dataset is necessary as the simulation might produce artificial noise or variability. This noise must be accounted for when the comparison with the distorted dataset is carried through, as this will include both the intentionally added distortion and the distortion added by the simulation. This enhances the comprehension of the patterns and relationships embedded within the data, along with a more accurate and robust analysis. Figure 4.1 illustrates the processing of the datasets.

This enhances the comprehension of the relationships and patterns embedded within the data, leading to a more precise and resilient analysis

4.2 GREY-NISLAB Keystroke Benchmark Dataset Syed

The GREYC-NISLAB Keystroke Benchmark Dataset dataset contains the collection of KD data from 110 volunteering individuals, whereas 70 users originated from France, using the AZERTY keyboard. 40 users originated from Norway and used the QWERTY keyboard. The participants were students, researchers, faculty members, administration staff, and others [Sye14]. Table 4.1 shows how the participants were distributed by age and gender, where the ages range from 15 to 65.

Age Group	Men	Women	Total
(30,65]	37	14	51
[15,30]	41	18	59
Total	78	32	110

Table 4.1: Users in the dataset separated by age group and gender, including totals.

All participants were to type 5 fixed, distinct phrases 20 times, 10 times with one hand and 10 times with both. The selected phrases were considered appropriate due to the familiarity of the names so that all participants, regardless of origin, could easily remember them. The phrases gradually increase in length, starting at 17 characters and up to 24, shown in Table 4.2, designed to capture and measure the participants' typing behavior across various difficulty levels.

Metadata such as gender, age, and handedness (left or right-handed) were collected. However, handedness will not be considered in this thesis; thus, the trials with only one hand have been removed, resulting in one participant writing the password ten times.

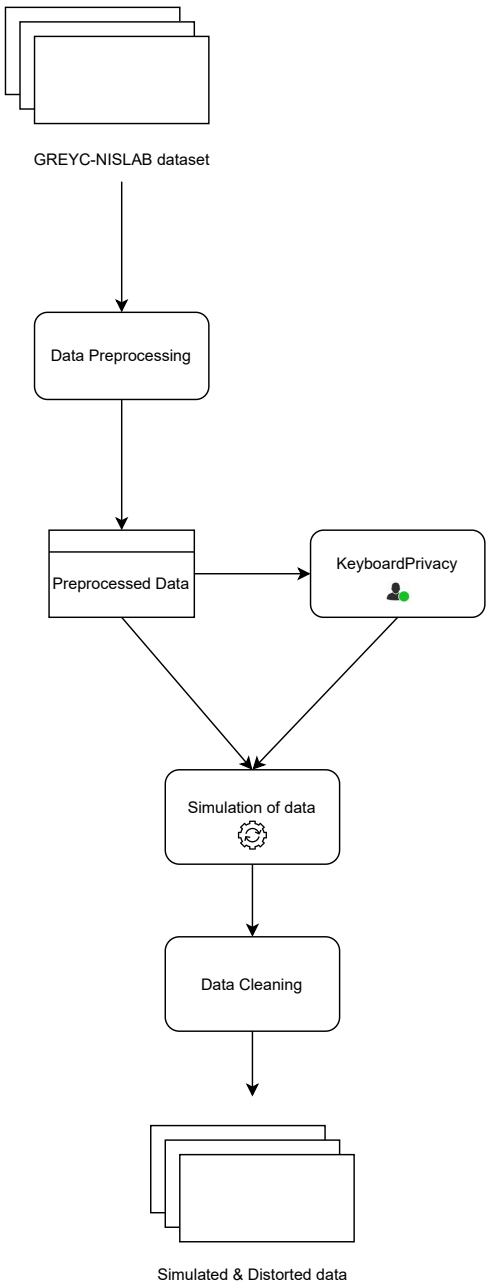


Figure 4.1: Flowchart of the simulation of data

Password	Size (char)
leonardo dicaprio	17
the rolling stones	18
michael schumacher	18
red hot chilli peppers	22
united states of america	24

Table 4.2: Passwords in the GREYC-NISLAB dataset

The dataset utilized is structured with the final format of:

- **ID:** Each participant received an unique ID.
- **Gender:** Gender of the participant
- **Age:** Age of the participant
- **Duration and Latencies:** The last column of the dataset contains the remaining $2n + 1$ values, where the first $n + 1$ values are duration values, followed by n latency values.

4.3 Keystroke simulation and collection

Several options are available for the simulation of keystrokes, such as **'pyautogui'** for Python [Al 19], **'java.awt.Robot'** for Java [Oraar]. Within this thesis's research, an existing simulation C++ program to simulate keystrokes was adopted. The C++ program was written by Tobias Moe and created for his Master's thesis with a similar purpose [Moe21b; Moe21c]. While the other alternatives mentioned are considered accessible and effective, personal programming experience introduces risks of error, making the already existing software a safer choice. In addition, by using an existing solution, a considerable amount of time is saved, making space for a more profound and comprehensive interpretation and analysis, aligning with the study's objectives. Some reformatting of the original datasets had to be done to fit the simulation.

To collect the data from the simulation, a minimalistic webpage was created, recording keystrokes in an input field. The keystrokes, and their associated timing information was collected with the use of jQuery in Javascript. This information was stored in a database using Flask, as this was more lightweight and suitable than Django[Moe21a].

The simulated data had a format of:

- **user_id:** Identifier of a user.
- **session_id:** Identifier for the session a user is currently typing in.
- **repetition:** The current repetition that the user is in.
- **type:** The type of action, either press or release of a key.
- **keycode:** The keycode of the key that was acted upon.
- **clocktime:** The clocktime of the action
- **lastkey:** The time since the last key was acted upon, counted in milliseconds.

For further analysis, the column *lastkey* is of interest. From this, both duration and latencies can be extracted, as well as other features like *average time of key press*, *average time of key release* and *max time*. For some unknown reason, neither *user_id*,*session_id*, or *repetition* was updated throughout the simulation. Therefore, this had to be manually added, in addition to the *age* and *gender*-attributres that were merged into the dataset, using a process analogous to the database join-operation.

4.3.1 Keyboard Privacy Plug-in

The Keyboard Privacy plug-in was enabled in two out of the three simulations. One simulation was done without, one simulation was done with the default settings, and the last simulation was done with a higher distortion. The default settings had a duration and latency delay of 200ms, while the last simulation was run with a duration and latency delay of 300ms. This way, with a variation in the amount of distortion, we can examine if the effects can be seen more clearly. This will help analyze whether the amount of delay added affects the ability to classify soft biometrics.

4.4 Data Cleaning and Missing Value Handling

After the simulated datasets had been collected, pre-processing and data cleaning procedures were carried out to ensure quality and suitability for further analysis. When utilizing machine learning algorithms, it is highly recommended that the underlying data is normalized beforehand, improving accuracy and reliability. Missing values, outliers, and inconsistencies were removed. This was achieved by using the *scale()* function in *R*. It subtracts the mean from the data, and divides it by the standard deviation. This ensures that the input is on a similar scale, and analysis on the same basis can take place.

Missing values that, for some reason, are not present or available in the original dataset are represented as -1000 or N/A, for example, due to a message left unfinished or sensor malfunctioning. These values are removed from the datasets. Negative durations were removed from the datasets, as this can not occur in a real-world context. A threshold for durations and latencies has been set at 500ms, as well as -500ms for latencies. This is done based on the assumption that typical typing behavior does not carry values above this point, even though there are possibilities for the latencies exceeding this for some key combinations. However, looking at simple, fixed texts, 500 ms is a reasonable value. These durations and latencies will be considered outliers and non-representative, causing the data to be removed.

Some inconsistencies during the simulation process were identified. It did not consistently capture all key presses and releases, and in some instances, it failed to capture the correct sequence of the press and release of the keystrokes. This led to incomplete data, and to mitigate the potential of this affecting the results, these instances had to be removed.

Chapter 5
Methodology

The objectives and research questions require extensive study, making it necessary to combine several types of research methodologies. To explore and determine if the proposed problem and its substantiating questions from the previous chapters are achievable, the methodology and steps taken are described in this chapter.

5.1 Literature Review

A literature review has been performed with the purpose of presenting the background and evolution of the classification within KD biometrics in Chapter 2 and 3. It aims to provide context for the work that will be carried out in this thesis and to do so by highlighting state-of-the-art research as well as popular methods currently used in the field. Additionally, the contribution of the work conducted in this thesis will be discussed.

5.2 Evaluation metrics

To assess the quality of predictions calculated by the machine learning models, different performance evaluation metrics are taken into use. These have a crucial role in the determination of the effectiveness of the models, as well as their suitability for particular tasks. The metrics provide a quantitative measure of performance from different perspectives and will typically involve a comparison of the predicted output and the actual output.

The choice of metric is dependent on the dataset that is under evaluation and the context for use. Accuracy is a common choice for classification evaluation but may not be a good choice for an imbalanced dataset. Here, a dataset may be biased and give the prediction of the majority class a high accuracy score while performing poorly on the minority class. Hence, the performance measure used must be nuanced and account for different error types.

These rates do not solely determine the percentage of effectiveness. They are dependent on several factors, including the quality of the data collected, the operational environment in which the system runs, as well as the performance of the algorithms used. Reduced quality of the biometric data will inevitably result in an elevated error rate, causing diminished effectiveness. Similarly, if the environment holds a high level of noise, the error rate may be affected.

5.2.1 Statistical Background and distance metrics

For the data analysis process, eight different distance metrics were utilized for comparing the distorted and the original timing data. The chosen distance metrics were included due to their common use within KD research and their ability to capture the impact of the distortion. It is worth noting that the primary objective in this study is not the evaluation and optimization of the performance metrics but rather capturing the differences between the different datasets and their characteristics. By including several distance metrics, a more comprehensive understanding of the similarities and dissimilarities from the datasets can be obtained, assisting further analysis and conclusions.

A brief explanation of the different distance metrics is provided to make the findings more clear and informative.

Arithmetic mean

The total number of observations is k, and each observation is noted x_i.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5.1}$$

Standard deviation

The total number of observations is k, each observation is noted x_i, and μ is the mean.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{5.2}$$

The Cosine Similarity

The similarity between two vectors is measured by the cosine angle between them [Kum20]. The cosine similarity between the vectors A and B is calculated by:

$$\text{Cosine Similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|} \tag{5.3}$$

Interquartile Range (IQR)

Interquartile Range (IQR) measures the spread or variability in a dataset and focuses on the middle 50% of the data. A particular advantage with IQR is that it is unphased by outliers [Fro21].

$$IQR = Q3 - Q1 \tag{5.4}$$

Where $Q1$ is the first quartile, the middle value of the smallest value and the median. $Q2$ is the second quartile, the median. $Q3$ is the third quartile, the middle value between the highest value and the median.

Euclidean distance

The Euclidean distance measures the similarity between two vectors or data points, by calculating the square root of the sum of the squared differences between the values [Kum20]. Equation 5.5 defines the distance metric of the two input vectors reference r and probe, p.

$$d_{ED}(r, p) = \sqrt{\sum_{i=1}^{n} (p_i - r_i)^2} \tag{5.5}$$

The Python function *numpy.linalg.norm* is used for the calculation of the Euclidean distance. Another alternative would be the *SciPy Euclidean* function; however, it exhibits a relatively slower performance[Moe21b].

Manhattan distance

The Manhattan distance measures the distance between two data points, determined by geometry in a grid-based system, often known as the *city block distance*. It is calculated by the sum of the absolute differences based on their coordinates in the Cartesian plane [Kum20]. The Manhattan distance is defined as;

$$d_{MD}(r, p) = \sum_{i=1}^{n} |p_i - r_i| \tag{5.6}$$

The Manhattan distance metric is implemented in Python, through the *numpy.linalg.norm* function, or with the *SciPy cityblock*-function.

Scaled Manhattan distance

The Scaled Manhattan distance is based on the Manhattan distance, where the absolute differences between the data points are divided by the absolute value of the data points[Bou12]. Scaled Manhattan distance can be defined as:

$$d_{SMD}(r,p) = \sum_{i=1}^{n} \frac{|p_i - r_i|}{a_i} \tag{5.7}$$

There is no direct Python function for Scaled Manhattan, thus, must be implemented manually.

Canberra Distance

The Canberra Distance measures the dissimilarity of two data points, by finding the absolute values of the difference, and each absolute difference is divided by the sum of the absolute values of the respective data points [Cud]. Canberra Distance is defined as:

$$d_{CaD}(r,p) = \sum_{i=1}^{n} \frac{|r_i - p_i|}{|r_i| + |p_i|} \tag{5.8}$$

The Canberra distance can be calculated by using the *SciPy canberra* function in Python.

Chebyshev Distance

The Chebyshev Distance calculates the greatest absolute differences between the values of the data points [Cud]. It can be defined as:

$$D_ChD(r,p) = \max_{i=1}^{n} |r_i - p_i| \tag{5.9}$$

The Chebyshev distance can be calculated by using the *numpi.linalg.norm*, or *SciPy chebyshev* function.

5.2.2 Confusion matrix

In order to compare the various models effectively, the metrics used to evaluate each technique must be used under the same conditions. Once the classification process has been carried out, the predictions will be assigned to one out of four categories represented in the Confusion matrix in Table 5.1. The Confusion Matrix consists of four entries that compare predicted outcomes with the actual outcomes; true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Ultimately, the accuracy illustrates the ratio between the correctly identified instances and total instances, as it can be seen in the downright corner in Table 5.1, and in percentage format in Equation 5.10.

		Predicted		
		Positive	*Negative*	
Actual	*Positive*	True Positive (TP)	False Negative (FN) **Type I error**	Recall $\frac{TP}{TP+FN}$
	Negative	False positive (FP) **Type II error**	True Negative (TN)	Specificty $\frac{TN}{TN+FP}$
		Precision $\frac{TP}{TP+FP}$	Negative Predictive Value $\frac{TN}{TN+FN}$	Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$

Table 5.1: Confusion Matrix containing the four different categories predictions can belong to. Illustration adapted from Chakravorty[Cha].

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{5.10}$$

Related to KD classification, the prediction of a user's typing is based on whether the user is an imposter or genuine. Similarly to the example, two possible errors can occur, type I - false negative and type II - false positive. The presence of two types of error results in a trade-off between achieving a higher precision or a higher recall. A high precision rate will typically have to sacrifice recall and will obtain a higher percentage of true positives, however, it will also produce an increased amount of false negatives.

5.2.3 Recall and Precision

The confusion matrix 5.1 shows us the equation for recall in the upper rightmost corner and the equation for precision in the lower leftmost corner.

Recall is the number of true positives out of the total of true positives and false negatives. It provides insight into the model's ability to identify the actual positive cases, such as classifying actual females as females.

$$Recall = \frac{TP}{TP + FN} \tag{5.11}$$

Precision is the number of true positives out of the total of true positives and false positives. It measures the validity of the model's predictions.

$$Precision = \frac{TP}{TP + FP} \tag{5.12}$$

5.2.4 F1-score

The F1-score measures the harmonic mean between precision and recall. It combines these values, providing a balanced measure to compare binary classification models and results.

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{5.13}$$

The F1-score ranges from 0 to 1, where the value of 0 indicates the worst balance of precision and recall, while a value of 1 indicates a perfect balance of precision and recall. It is appropriate for use when the classes to be examined are imbalanced, as it takes into account the cost of false positives and false negatives. When using accuracy on imbalanced data, the classifier of the majority of the class will always obtain higher accuracy, even when performing poorer in the minority class, making F1-score more reliable in these cases.

5.3 Feature selection

The selection of features provides the base of the learning of the machine. A subset of relevant data is extracted and utilized. The most relevant features in KD are latencies and durations, hence used for this study. Initially, only these features were planned to be utilized, however, due to the data cleaning process removing a portion of samples, a more effective solution ended up with using the averages of the key presses and key releases, the averages of durations and average latencies. Additional features extracted were maximum press time, and count of key rollover, i.e., the amount of subsequent key presses before the key releases.

80% of the features and corresponding labels were used for training, and 20 % were used for testing, a normal separation for machine learning. This ensures adequate learning, allowing the machine to capture the underlying structures and avoid overfitting.

The process and architecture of the methodology is illustrated in Figure 5.1, and as mentioned, the process of simulation is illustrated in Figure 4.1.

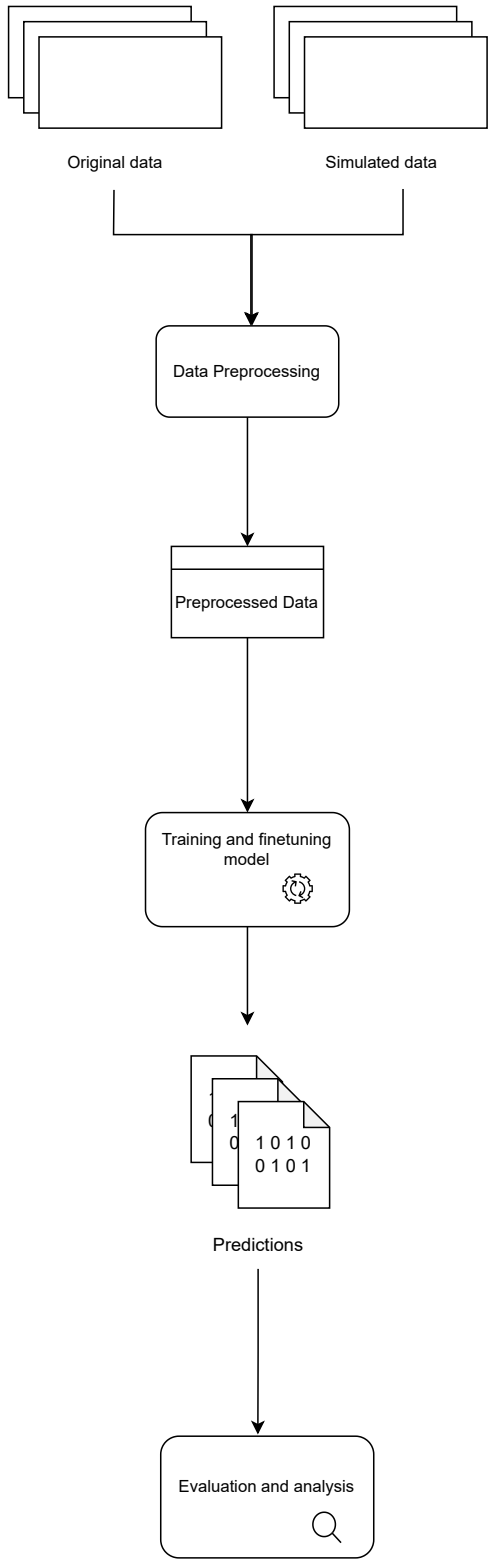


Figure 5.1: Flowchart of the classification process

Chapter 6
Results

This chapter presents the statistical results and the machine learning model results. First, the comparison between the original, simulated, and distorted datasets is provided. Furthermore, the results from the classification of soft biometrics by machine learning are provided and presented. Throughout these chapters, the passwords will be referred to as PWx, where x is the number of the specific password, and the distorted passwords will be referred to as Dx, specifically $D1$ and $D2$. All measurements are in milliseconds.

6.1 The effect of simulation

6.1.1 Statistical results

The mean of durations and latencies, including the standard deviation, for different soft biometric groups, are listed in table 6.1. The data has undergone a data-cleaning process, ensuring no outliers influence the results. Referring to the users in the "*gender*"-category in the original dataset, we find that the average duration and latency for males are, respectively, $83.41ms$ and $115.56ms$, whereas women types somewhat slower, $90.03ms$ and $114.59ms$.

When dividing the users into age groups, we find that the duration and latency for age group < 30 are respectively $86.02ms$ and 105.16, while for the age group > 30, the values are respectively $84.68ms$ and $125.20ms$.

In the simulated dataset, all the values are slightly added some delay, where it is found that males obtain a mean duration of $89.91ms$ and a mean latency of $118.79ms$. Females obtain a mean duration of $100.29ms$ and a mean latency of $122.53ms$. The age group < 30 shows a mean duration of $90.95ms$ and a mean latency of $107.19ms$. The age group > 30 obtain a mean duration of $94.72ms$, and a mean latency of $132.66ms$.

		Durations		Latencies	
		Mean	**SD**	**Mean**	**SD**
Original dataset	**Male**	83.41	29.16	115.56	89.66
	Female	90.03	35.62	114.59	94.29
	Under 30	86.02	31.78	105.16	88.08
	Over 30	84.68	30.91	125.20	92.67
Simulated dataset	**Male**	89.91	38.36	118.79	95.47
	Female	100.29	40.21	122.53	97.85
	Under 30	90.95	42.15	107.19	94.42
	Over 30	94.72	25.82	132.66	96.31

Table 6.1: Table of statistical measurements of the passwords combined

6.1.2 Interquartile Range Comparison

Table 6.2 presents the summary of the statistical dispersion in the form of IQR measurement, subtracting the first quartile (25%) of the data from the third quartile of data (75%). The mean IQR of the simulated datasets is 62.2, meaning that 50% of the data falls around the range of $61ms$ and $66ms$. For both $D1 - IQR$ and $D2 - IQR$, the mean is relatively higher, respectively, $98.15ms$ and $88.2ms$, a significant shift upwards from the distorted data. Figure 6.1 and Figure 6.2 illustrates the comparison visually with a graph and a box plot. The box plot emphasize the different value ranges.

Dataset	**Simulated IQR**	**Distorted1 IQR**	**Distorted2 IQR**
PW1	66	96	67
PW2	54	105	69
PW3	69	98	106
PW4	61	95	99
PW5	61	96.75	100
Mean	62.2	98.15	88.2
SD	5.72	3.98	18.65

Table 6.2: Comparison of IQR Values

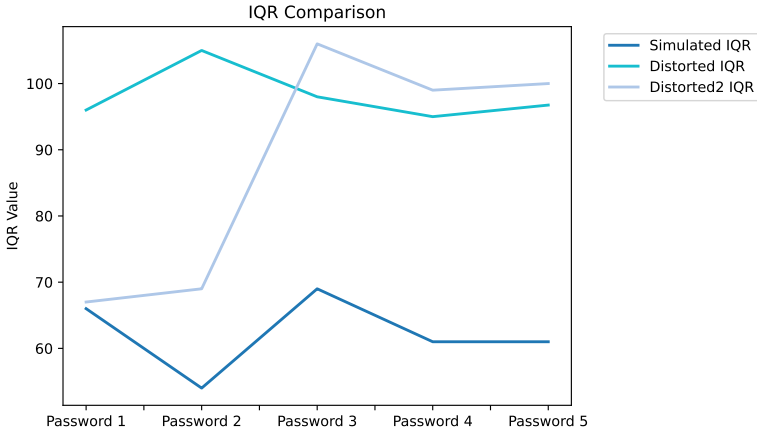


Figure 6.1: Interquartile range comparison

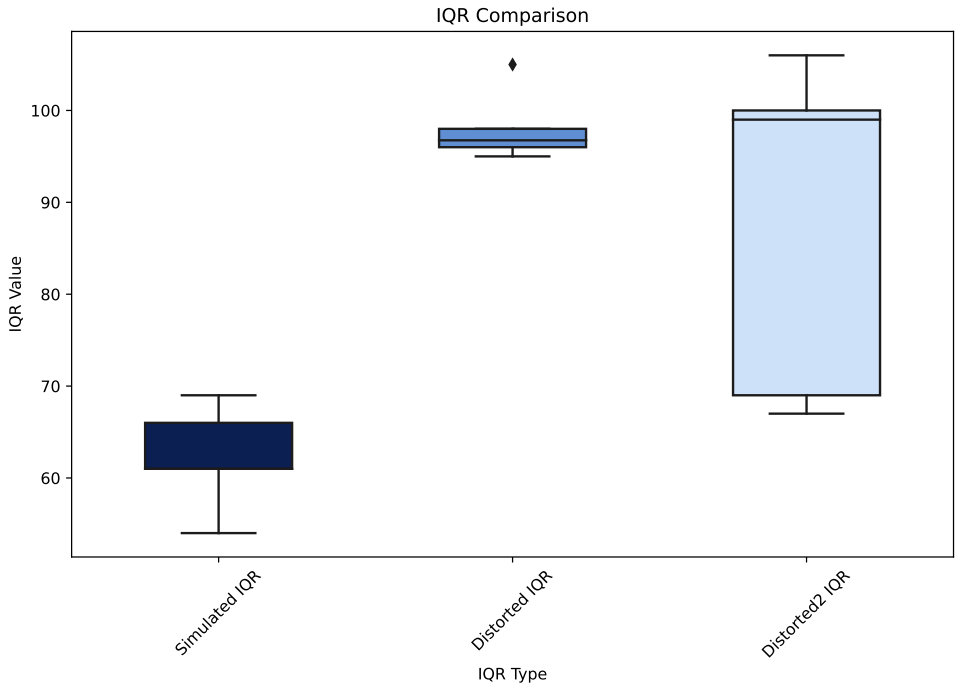


Figure 6.2: Interquartile Range comparison

6.1.3 Cosine Similarity

Table 6.3 provides a side-by-side comparison of the cosine similarity between the two distorted and the simulated dataset. The Cosinue Similarity expresses the similarity between the datasets, with values closer to 1 meaning nearly alike, while values closer to 0 express high dissimilarity. The mean cosine similarity of $D1$ and the simulated dataset is 0.858, where the lowest and the highest similarity is, respectively, 0.760 and 0.908. The mean cosine similarity of $D2$ and the simulated dataset is somewhat lower than $D1$, with a value of 0.754, where the values range from 0.648 and 0.995.

Dataset	Cosine Similarity D1	Cosine Similarity D2
PW1	0.821	0.995
PW2	0.908	0.933
PW3	0.988	0.405
PW4	0.760	0.790
PW5	0.814	0.648
Mean	0.858	0.754
SD	0.089	0.237

Table 6.3: Cosine Similarity of the simulated and distorted datasets

6.1.4 Distance comparison between the datasets

The distorted datasets' distances from the simulated dataset have been calculated and presented in Table 6.4. The distances presented are Euclidean, Manhattan, Scaled Manhattan, Canberra, and Chebyshev distances, and used to provide a quantitive measure of similarity of the datasets. A heatmap of each distorted password and their corresponding values can be found in Figure 6.3 and Figure 6.4. The heatmap provides an illustration of the similarities and dissimilarities. The color palette represents the magnitude of the distances, where a lighter shade of blue represents a higher grade of similarity, and a darker shade of blue represents a higher grade of dissimilarity.

For $D1$, the Euclidean distance values range from $14.06ms$ to $24.86ms$, with a mean of $18.86ms$ and a standard deviation of 3.24. The span is slightly smaller for $D2$, which ranges from $13.90ms$ to $22.11ms$, with a mean of 17.89 and a standard deviation of $42.94ms$. The Manhattan distance in $D1$ ranges from 115.55 to 254.23, while in $D2$, it ranges from $93.85ms$ to $196.34ms$. The Scaled Manhattan distance in $D1$ has a mean of $57.76ms$ and a standard deviation of $10.89ms$, and ranges from $40.90ms$ to $69.29ms$, while in $D2$, it ranges from $37.21ms$ to $85.73ms$, obtaining a mean and standard deviation of respectively $56.68ms$ and $|5.99ms$. The table yields that $PW5$ in $D1$ generally has the highest distance value, while in $D2$, $PW2$

generally produces the largest distances. On the other side of the scale, we find that $D1 : PW3$ and $D2 : PW3$ have the lowest distance score.

		Euclidean	Manhattan	S. Manhattan	Canberra	Chebyshev
D1	**PW1**	17.05	143.94	45.88	18.05	5.06
	PW2	18.66	166.49	62.54	21.72	5.42
	PW3	14.06	115.55	40.90	14.09	4.45
	PW4	19.68	189.83	69.29	23.50	5.51
	PW5	24.86	254.23	69.17	21.29	6.39
	Mean	18.86	174.01	57.56	23.46	5.37
	SD	3.24	42.94	10.89	3.34	0.59
D2	**PW1**	15.89	148.14	52.26	20.36	4.87
	PW2	19.46	196.34	85.73	25.52	5.48
	PW3	13.90	93.85	37.21	10.18	4.90
	PW4	18.10	147.81	57.49	15.05	5.69
	PW5	22.11	173.57	50.73	17.77	6.62
	Mean	17.89	151.94	56.68	17.77	5.51
	SD	2.84	34.19	15.99	5.13	0.64

Table 6.4: Summary of the distance statistics performed on the simulated and distorted passwords

6.2 Classification results

The different datasets are utilizing a 80% training and 20 % testing split. The model performed with a radial basis kernel function, a cost of $C = 100$, $\gamma = 2$ *and* $\epsilon = 0.1$. *For both cases, the features of average duration and average latencies were utilized, and for enhanced performance and increased balance, max press time and a count of how many times users executed a key rollover were included.*

6.2.1 Gender classification

Table 6.5 showcases the results of the classification of gender. The confusion matrix "calculated" from the simulated dataset with no distortion is presented in Figure 6.5a. This can be interpreted as 67 true positives, showing that these are correctly identified females, and 16 false positives, male users incorrectly identified as females. There were 62 true negatives, correctly identified males, while 11 false negatives, falsely identified as males. This leads to the performance metrics in Table 6.5, where the attained precision, correctly predicted positive (female) cases out of all cases, was 80.7%. The recall, the proportion of actual positive cases that were correctly classified, was 85.9%, and the harmonic mean, F1-score, was 0.832. Recall that the

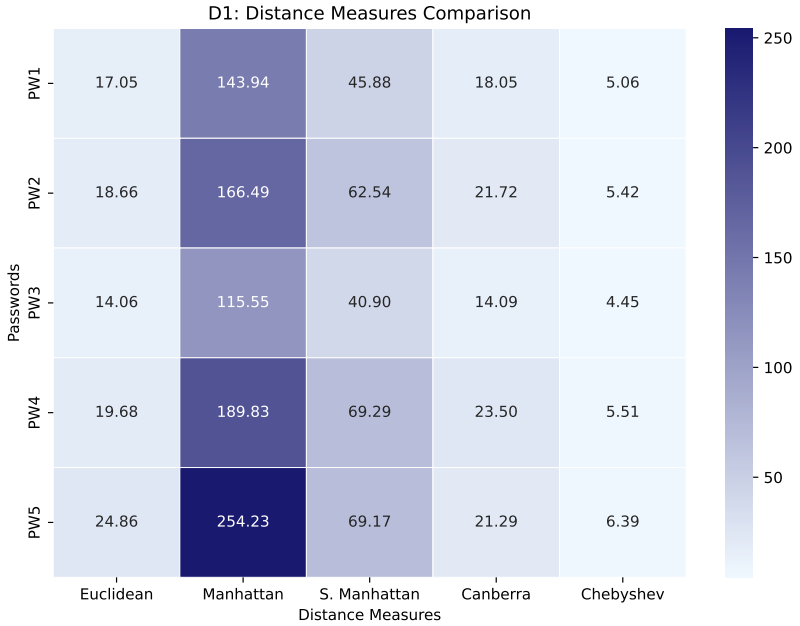


Figure 6.3: Heatmap of Distance Metrics with D1

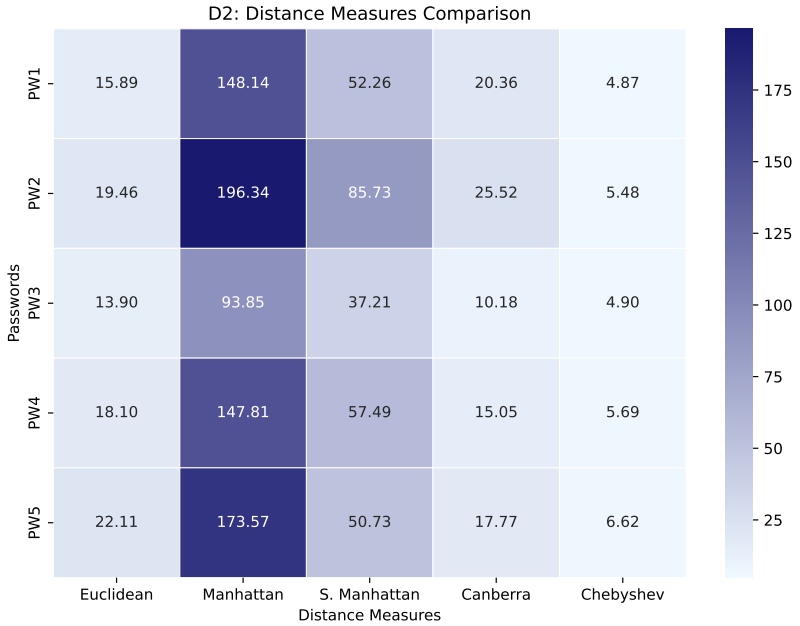


Figure 6.4: Heatmap of Distance Metrics with D2

F1-score represents the trade-off between the precision and the recall of the model. The overall accuracy of the model achieved was 82.59%, implying that our model predicted the correct instances approx. 83% of the time.

The confusion matrix of the distorted dataset with $0 - 200ms$ distortion is presented in 6.5b, consisting of 69 true positives, 61 true negatives, 7 false negatives, and 18 false positives. Referring to Table 6.5, a precision of 79.3% and a recall of 90.7% was obtained, leading to an F1-score of 84.7%. The overall accuracy attained was approx. 84%, achieving a higher performance across all classes compared to the non-distorted dataset.

The second distorted dataset, with a randomized delay of $0 - 300ms$, obtained the confusion matrix presented in 6.5c. There were observed 66 true positives, 54 true negatives, while the false negatives and positives resulted in, respectively, 12 and 24. Of all the predicted positive cases, 73.3% was correctly identified (precision). The recall rate demonstrated was 84.6%, producing a recall rate of 78.6%. The overall accuracy was 76.92%, meaning the gender of the user was predicted correctly approx. 77% of the time, somewhat lower than the simulated and the Distorted 1 datasets.

	Precision	**Recall**	**F1-score**	**Accuracy**
Simulated	0.807	0.859	0.832	82.6%
Distorted 1	0.793	0.907	0.847	83.87%
Distorted 2	0.733	0.846	0.786	76.92%

Table 6.5: Results of classification on gender on all the passwords with 80% Training, 20% testing data

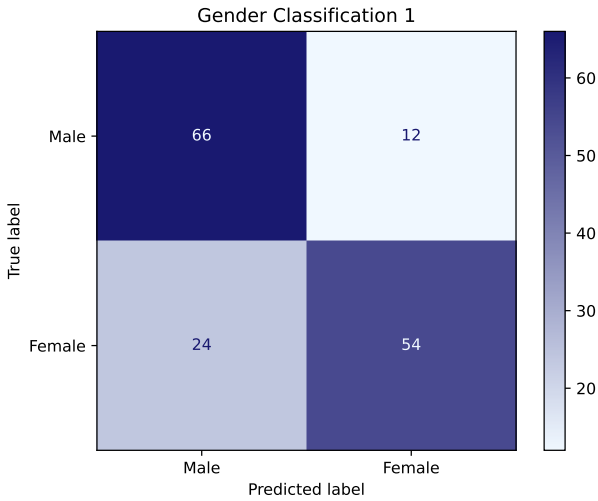
Grid Search

Grid search for gender classification was performed, resulting in best results for $Cost = 100$, gamma $\gamma = 2$, and epsilon $\epsilon = 0.1$

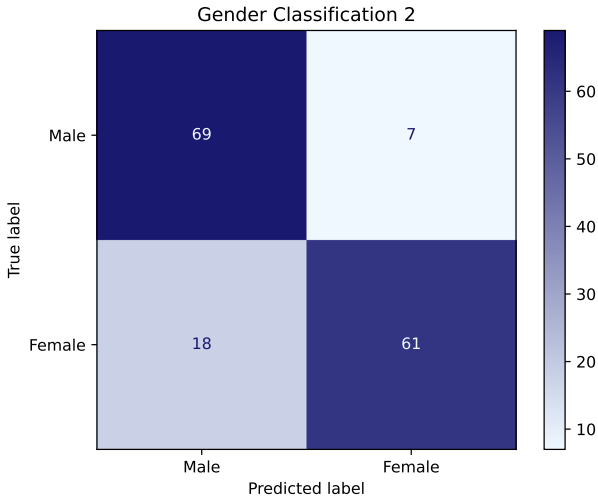
6.2.2 Classification on age

The SVM classification of age achieved worse performance than the model for the classification on age. The results will be further discussed in the next chapter.

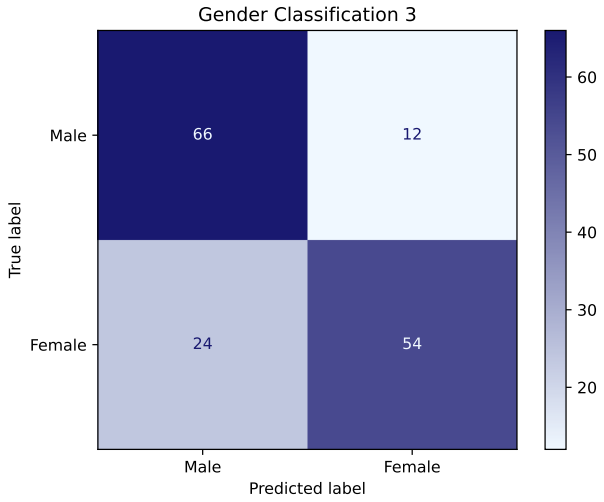
The simulated dataset achieved 63 true positives and 50 true negatives, 20 false negatives, and 23 false positives, represented in the confusion matrix in Figure 6.6a. As noted in Table 6.6, the F1-score reached 74.5%, the balance of the precision of 73.2%, and recall of 75.9%. Ultimately, the overall performance of the model in terms of accuracy was 72.43%.



(a) Confusion Matrix of gender classification on all the passwords with no distortion



(b) Confusion Matrix of gender classification on all the passwords with general settings of 200ms distortion in duration and latency



(c) Confusion Matrix of gender classification on all the passwords with settings of 300ms distortion in duration and latency

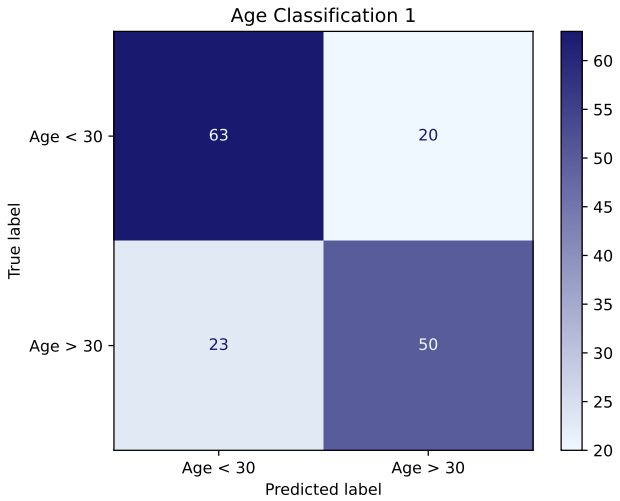
Figure 6.5: Confusion Matrices for gender classification on different settings

The Distorted 1 dataset reduced the overall accuracy to 59.63%. The confusion matrix in Figure 6.6b represents the true positives of 37 and the false negatives of 16. Furthermore, the false positives and true negatives were both equal to 28. Table 6.6 reveals an increase of misclassifications with distorted data, where precision is reduced to 0.569, recall is reduced to 0.698, and the F1-score is reduced to 0.627.

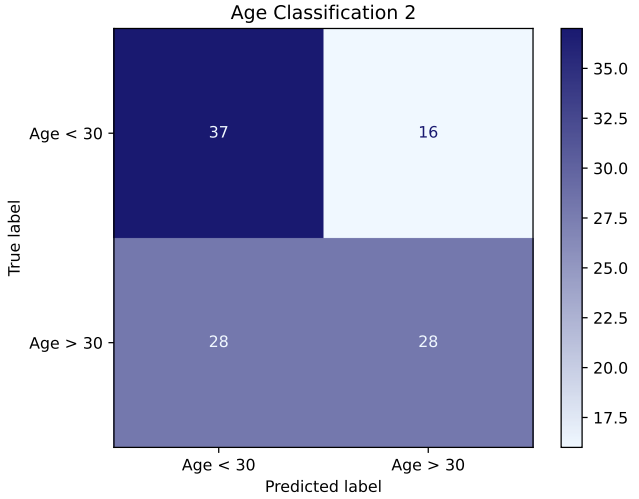
Lastly, the Distorted 2 dataset, had a slight improvement from $D1$ in overall accuracy with a score of 65.4%, also noted in Table 6.6. The Confusion Matrix in Table 6.6 presents the recall value of 0.614, a precision value of 0.698, and an F1-score of 0.654. This is derived from 6.6c, where the model predicted 55 true positives, 32 false negatives, 22 false positives, and 51 true negatives.

	Precision	**Recall**	**F1-score**	**Accuracy**
PW1 - Simulated	0.732	0.759	0.745	72.43%
PW1 - Distorted 1	0.569	0.698	0.627	59.63%
PW1 - Distorted 2	0.698	0.614	0.654	65.4%

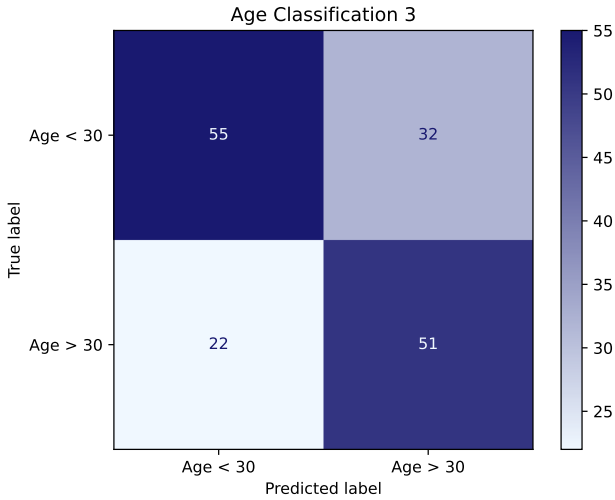
Table 6.6: Results of classification on age on all passwords, with 80% Training, 20% testing data



(a) Confusion Matrix of age classification on all the passwords with no distortion



(b) Confusion Matrix of gender classification on all the passwords with general settings of 200ms distortion in duration and latency



(c) Confusion Matrix of age classification on all the passwords with settings of 300ms distortion in duration and latency

Figure 6.6: Confusion Matrices for age classification on different settings

Grid Search

All models for age classification obtained a cost of $c = 100$, gamma $\gamma = 2$, and $\epsilon = 0.1$. However, the distorted 1-dataset achieved better results with $cost = 1$, hence, the setting was employed for this dataset.

Chapter 7
Discussion

This chapter interprets and discusses the results that were presented in the earlier chapter, with the intention of answering the research question: "Is it possible to detect soft biometrics, such as age and gender, even when the KD data is distorted?"

7.1 The effect of simulation

The simulation of the datasets is necessary to create a reference point for the analysis. It is to be expected that the simulation itself will produce some noise or delays, as identical timing values cannot be achieved. One must keep this added noise in mind when reading these results to attain a valid and reliable analysis. From the results in Table 6.1, we can confirm that the simulation itself produces some delay in the durations and latencies, where the mean value and standard deviation have slightly increased in all measures. The noise can be due to the C++ simulation program, which implements a sleep function that might be set off before or after the initial set value. This causes the delay in the keystrokes. In addition, the capturing of the keystroke timing can be affected by the JavaScript functions on the webpage. The amount of delay added in the durations ranges from approximately $4 - 10ms$, and for the latencies, the range is approximately $2 - 8ms$. It does not constitute a significant portion of the keystroke timings, however, it is a notable proportion, hence, it must be kept in mind while examining this work.

7.2 Comparison of simulated and distorted data

Interquartile Range (IQR)

Recall that D1 was added random values from 0-200ms in 50% of the samples, while D2 was added random values from 0-300ms in 50% of the samples. The range of where 50% of the timing values in the dataset land, is shown through the measuring of the interquartile range of the durations, showing that there is a significant upward shift in the distorted datasets, shown in Table 6.2 and visually in the boxplot 6.2.

When comparing the simulated IQR with the D1 IQR, we find that the simulated dataset has a broader spread, implying that the simulated dataset exhibits a higher variability. This could be due to the noise added from the simulation process, or that the distortion in D1 reflects the specific nature of the added distortion, which could have narrowed down the range of value.

When comparing the distorted IQRs, we find that the IQR of D1, that range from 95 to 98, is smaller than that of D2, ranging from 67 to 100. This indicates that the distortion with D2 has a larger spread than the distortion in D1, suggesting a greater measurement error or a greater variability and range of values with a lower distortion setting. This is consistent with both the delay added from the distortion plug-in, as well as the simulation delay. This statement is further supported by the standard deviation, indicating that the data points in D2 are more dispersed around the mean compared to D1.

In relation to user behavior analysis, wider IQR observed can potentially lead to more diverse patterns and inconsistencies in the features, posing a challenge for further analysis. Additionally, a larger standard deviation in D2 indicates a greater dispersion of keystroke timing information that may require considerations in the data pre-processing, feature extraction and model.

It is important to recognize that even if it is possible that higher delay values can be added to D2, the delay is randomized, meaning that the data in D2 can exhibit less distortion than the data in D1. This observation holds true for both Password 1 (PW1) and Password 2 (PW2) and the mean IQRs, as evidenced by the data presented in Figure 6.1.

Ultimately, there are distinctive differences between the distorted datasets, and the simulated one, hence, using IQR could help indicate if distortion have taken place. However, it would be difficult to determine whether the data is distorted if the simulated dataset's IQR values where removed from the figure, which would be the case in real-world scenarios.

Cosine Similarity

The cosine similarity, expressing the similarity of the datasets, is relatively high, implying that the datasets are all close to being parallel, or might even be parallel in a higher dimensional space. The similarity between the simulated and the D1 dataset is averaging 0.858, which indicates that the distortion in the bigger picture has not made a significant impact. The cosine similarity of D2 is somewhat lower, with a value of 0.754, which implies that there is more distortion added. This was to be expected, as the range of values added is higher. The similarity value is still relatively high, with 75% of the values corresponding to the simulated ones. With a

higher standard deviation for D2 than for D1, respectively 0.237 and 0.089, a more consistent similarity with D1 can be suggested. The keystroke timing information in D1 has a higher consistency, and lower variability with the simulated dataset, indicating a higher degree of similarity.

The cosine similarity results, combined with the information on the range of added values, provide valuable insights into the impact of different distortion levels on KD analysis. In particular, the additional details regarding the range of added values in "D1" and "D2" shed light on the extent of distortion and its implications for the accuracy and reliability of KD systems.

There is a considerable amount of difference in the range of the IQR values, implying that there might have been some manipulation of the timing information. Furthermore, the cosine similarities show that there remain fundamental similarities, making it possible to extract important characteristics from it. It is essential to keep in mind that all differences can be due to other underlying factors of the user, as mentioned in Chapter 2, such as emotions, change of environment, familiarity and skills with the keyboard, as well as natural variability. Our typing patterns are not solely based on static factors but are influenced by unforeseen and continuous changes. However, the IQR implies that there is a consistent pattern of change, suggesting that there is potential for influencing factors beyond natural.

These results may not be enough evidence to prove that distortion has happened directly. However, it has the potential of implying or suggesting it. Combining the IQR and the cosine similarity results, the impact of different levels of distortions on KD can be further examined. As mentioned, when considering if a dataset is distorted, there is no knowledge about the current user, so the decision must be taken relative to the expected behavior and the deviation from it. The challenge, in this case, would be to decide what is the expected behavior. If the chat forum requires the user to disclose personal information, such as age and gender, to access the service, expected behavior can be determined. It can be assumed that a chat service for children would have policy requiring this, but if this is not the case, it can be challenging to establish a reference behavior. It is also possible to state false information, causing the reference table to be skewed.

7.2.1 Metric Distances between the datasets

One approach to investigating the impacts of distortion can be through distance metrics. Recall the results from the distance metrics (explained in Chapter 5.2.1 calculated between the distorted datasets and the simulated dataset from Table 6.4. Gaining an understanding of the underlying metrics can yield valuable insights regarding their optimal usage in a variety of contexts. This, in turn, serves to enhance our overall comprehension of distortion.

Recall that the Euclidean and Manhattan distance metrics employ geometric and grid-based calculations to determine the shortest path between points [Kum20; Sha20]. A comparison between the distortion in D1 and D2 reveals that D1 displays a greater degree of distortion. This observation suggests that the overall distortion, when considered as a whole, can potentially exert a greater influence than individual distortions, even if some of the latter are extreme.

Recall that Scaled Manhattan distance metric standardizes the variations between data points. The findings indicate that the two datasets produced relatively similar results, implying that the degree of distortion relative to the original data is similar in both datasets, despite any discrepancies in the actual values. This shows how distortion behaves in this context.

Although the Canberra distance metric has a lower distortion range, it still displayed a higher average distance for D1. This suggests that D1 has more significant relative differences because of distortions compared to D2. Furthermore, points with very low or high Canberra distances are likely to reflect distortions in the dataset, emphasizing the importance of considering distance metrics for revealing nuanced information.

The Chebyshev distance metric is used to measure the maximum coordinate difference. This metric showed a slightly higher mean for D2, which can be considered consistent with expectations as D2 has a larger range of possible added values (0-300ms). This implies that Chebyshev is good at detecting the effects of severe distortions, and is especially helpful when the maximum difference is more important than the cumulative or average differences.

7.3 Soft Biometrics Classification

7.3.1 Choice of Machine Learning Classifier

After the initial literature review, the choice of machine learning model fell on SVM for several reasons. Firstly, the SVM is a well-established and reliable choice in the field of classification problems. The SVM has the ability to handle high-dimensional data, like KD data, and find the optimal hyperplane to separate it. Being robust and effective, with a solid mathematical foundation, it is a widely respected choice amongst researchers in a wide range of fields.

Secondly, SVM is also a frequent choice amongst studies in KD, providing a substantial foundation for this work, and a demonstration of viability in this context. Following the established practices in this field of research, building upon the existing findings seemed like a reasonable option. A former Master's thesis by Tobias Moe

specifically mentioned SVM as a machine learning classifier to be examined, which played a big role in this decision [Moe21b].

However, throughout the project, the thought of using unsupervised machine-learning approaches seemed interesting. A real-time application would collect unforeseen and unknown data, which is the advantage of unsupervised learning models. Their flexibility to adapt to new data could be a significant benefit in situations like real-time scenarios, where the model must learn from unseen data.

The initial choice of utilizing SVM has provided valuable insights and a new perspective on the complexity of the matter and may be a guide for further research directions in KD.

7.3.2 Gender classification

Imbalanced dataset

The datasets utilized in this study lack balance in the gender class. The class of males is significantly overrepresented, which may lead to a bias toward this majority class during training and testing of the machine, and the classification performance will then end up suboptimal.

To account for this, the minority class of females was artificially inflated, meaning that new samples in this class were produced from the already existing ones. This balance out the dataset, and can improve the performance of the classification. However, creating new samples from already limited data might result in augmented data, the model might only recognize these specific instances, overfitting, and limiting the ability to generalize new, unseen data. Recall that overfitting is when the model learns the data too well, described in Section 2.5.5.

Alternatively, the majority class can be undersampled, where the majority class will have random samples removed until a balanced ratio is achieved. However, due to the already low quantity of data, this risks losing potentially important information found in the samples. Furthermore, choosing different evaluation metrics might be of interest, as accuracy might not be reliable when dealing with imbalanced datasets and biased classifiers. Metrics like precision, recall, and f1-scores embrace both false positives and negatives, providing a more realistic assessment.

Interpretation of results

A machine learning model is trained on specific data, with the aim to learn unique patterns that distinguish the classes and make classification possible, in this case, male and female. When introducing noise, such as distortion of input data, it is expected to have an impact on these patterns, making the classification process more difficult. Thus, a lowered performance of the machine would be a logical assumption, as the underlying patterns might be harder to detect.

The obtained results suggest otherwise, where both the D1 and D2dataset achieves higher accuracy than the simulated one. As accuracy can be a biased metric, precision, f1-score, and recall must be evaluated. However, also these exhibit better performance than the simulated dataset. These results imply that the model is better at dividing the males and females into the correct class when distortion is present.

Multiple explanations can be a possible cause for this. One explanation is that the distortion added to the timing information has amplified the underlying patterns. This way, features that may have had a subtle effect are now accentuated, and the hyperplane can maximize the surface further than earlier. In addition, the risk of overfitting due to excessive training on outliers and noise might have been mitigated when distortion is present. Thus, facilitating a more accurate classification of the samples into gender classes. A further possible approach could suggest that there exists some type of noise or bias in the original dataset, and the distortion added has unintentionally reduced this noise or bias, making the SVM perform more accurately.

7.3.3 Age Classification

When classifying on age group, it is reasonable to believe that children type slower than adults. This could be due to the continuous exposure and consistent interaction with technology and keyboard-based devices amongst adults. Adults have the need for keyboard device usage in both personal and professional contexts, facilitating more familiarity, muscle memory, and motor skills with the keyboards. Thus, a more efficient behavior with the keyboard. On the other hand, children might not be as exposed to these devices, as well as hand-eye coordination or motor skills are not as developed as an adult, resulting in a lower average typing speed. It is important to note that the use of technology from an early age is increasing, which has the potential to accelerate the learning process of typing on the keyboard, thus increasing the typing speed. Similarly, some adults lack access to and interaction with keyboards, lowering the average typing speed.

The dataset is diving the age groups into two; under 30 years old, and over 30 years old. This division is in itself quite broad. When referring to children, it is reasonable to assume that ages under 18 or 20 are in discussion. In addition, after

the age of 20, many individuals enter the workforce or start their education, exposing them to keyboards, thus enhancing their typing skills and increasing their typing speed. Some elderly individuals have, on the other hand, had no exposure to digital keyboards throughout their lives, leading to a lower typing speed. This leads to high variability within the same age groups, as individuals might have severely different typing patterns. Thus, categorizing age into binary categories can be argued not to be the most efficient approach. However, the dataset is limited, it contains too few participants of ages under 20, hence, it would not be possible to obtain reliable results.

The obtained results show consistently relatively low values in accuracy, precision, recall, and f1-score. The metric performance can be due to the high complexity and high-dimensional nature of the KD data, which results in difficulty in the classification of age. This corresponds with existing literature, showing that age classification has achieved lower performance with the same dataset [Sye14]. The performance is slightly increased with the use of average duration and latency features, and even further enhanced with the distortions. As previously assumed, the introduction of distortion would predict the performance to worsen. However, as with the gender category, this was not the case. The overall performance indicates that the SVM is struggling with both non-distorted and distorted data, but there is some potential to classify age with a high degree of uncertainty. The feasibility increases slightly with a distorted dataset.

7.3.4 Difference in performance between age and gender classification

The variabilities and complexities differ between the age groups, which has an impact on the performance of the classification task. The results imply that the correlation between KD and gender is higher than between KD and age. This may be due to KD embodying underlying patterns and features that align with gender to a higher degree than with age. Physical attributes such as strength, agility, and finger size might be underlying factors for the differences.

Furthermore, the variabilities and complexities in the data appear to be more prominent in the age category, leading to a more challenging classification. Age covers a broad range of variables, continuously changing, and the variability can turn out more complex and subtle. Within the same age group can also exhibit significant differences in typing patterns, based on cognitive abilities, familiarity with devices, motor skills, and environment, to name a few. The findings suggest that the impact of distorted data on age and gender differs, demanding a careful consideration of classification models and hyperparameters. Features that may be effective for gender classification might not be as effective for age classification. The gender-specific

patterns might appear more distinct, however, societal and cultural factors might increase the complexity of this category.

7.3.5 Possible factors for same or better performance with distorted data

Existing literature within the field of KD and classification has primarily been performed on data without distortion. An intuitive perception is that the distortion will hide the characteristics and patterns, making it harder to utilize and analyze the data. Surprisingly, the findings of this study reveal that distortion can in certain contexts enhance the performance of machine learning models, particularly Support Vector Machines (SVM). This counter-intuitive outcome casts doubt on this bias against distortion.

A possible explanation for these results lies in the distortion's ability to amplify the existing patterns in the data, facilitating the model's capacity to distinguish complex and subtle variations. Consequently, this enhanced distinctiveness can possibly lead to an improvement in the model's ability to correctly classify the data.

Furthermore, existing biases in the data can possibly be mitigated by the addition of distortion. Outliers and extreme values might be balanced out, providing a more balanced and impartial representation of data. Hence, a more objective and reliable classification can take place.

7.4 Hyperparameter tuning

Hyperparameter tuning in KD is a complex task due to the variability of features which leads to a higher dimensional feature space. Individuals have unique, dynamic typing patterns, making it difficult to generalize the features. The case of data imbalance in the gender category also makes the tuning of hyperparameters more challenging, as the model can be biased towards the majority class. It shows that there is a need for different hyperparameters for the original dataset and the distorted dataset.

7.5 Limitations and weaknesses of the research

Both external circumstances and personal oversights have led to limitations regarding this study. It is important to have in mind that the evaluation metrics represent a simplified view of the model's performance. Class imbalance, dataset size, distortion type and degree, and type of SVM kernel can have a significant impact on the results. Thus, the generalizability of the model in different contexts is a crucial aspect to consider.

The main concern has been the amount and quality of data. The amount of data may be considered insufficient when training a machine learning algorithm for the classification of such complex classes as male and female. Risks of overfitting, sampling bias, and guessing instead of genuinely learning and predicting. The oversampling poses a risk of overfitting during gender classification. The model's ability to predict the minority class is not always good enough to be reliable. Overfitting might be illustrated through the low recall and f1-scores, hence a risk in the age classification.

The study has been rather intensive and demanding regarding computational resources and time, resulting in a significant expenditure of time. Several setbacks were encountered during the simulations and had to be restarted multiple times, and was more time-consuming than originally expected. The simulation's loss of entries would seem to lead to consequential error, where the difficulty of the acquisition of the wanted features increased, and had to be exchanged with other features.

In the context of this analysis, it is essential to remember that behavioral biometrics is dynamic, and there is individual variability. It is affected by the external environment as well as correlating with emotions, personality, and skills, which is challenging to quantify. Thus, the human aspects must be taken into consideration throughout KD research.

Chapter 8
Conclusion

This chapter concludes the work undertaken in this thesis in relation to the goal and research questions introduced in Chapter 1. Following, contributions to the KD field from this work will be presented, before ultimately proposing potential topics to further extend and build upon this research.

8.1 Conclusion

The study was executed by utilizing a simulation program, producing three new datasets, where two were intentionally introduced with varying amounts of delay. Similarity and distance measures were calculated, and the results indicated that a sufficient level of similarity was present to be able to classify the characteristics. Simultaneously, it was revealed that enough distance from the distorted to the non-distorted data was present to suspect distortion.

The primary concern for the study of KD classification with distorted data has been the amount of data, which initially was limited. This was further reduced due to losses encountered during simulation. Furthermore, the gender classes were imbalanced. To tackle this, oversampling was performed for the minority class, and different feature combinations were explored. This resulted in the best performance achieved with the combination of average durations and average latencies. Hence, this research has a notable contribution to the identification of soft biometrics with distorted data, with the use of KD.

The following present a conclusion of the work in the context of the sub-questions and the overall goal:

Sub-Question 1: *What difference do distorted and non-distorted timing data have when it comes to performance?*

The findings from the analysis reveal that KD distortion can, in certain contexts, enhance the performance of the SVM. Specifically, it enhanced in accuracy, recall, precision and F1-score, suggesting that the addition of delay in KD data can contribute to an improved classification outcome. The reasons for this could potentially be explained by bias mitigation, noise reduction, or enhanced differentiation. However, it is essential to note that the results depend on the nature of the data used, the distortion, and the hyperparameters in the SVM.

Sub-Question 2: *How should distorted data be handled?*

The best results for the distorted datasets were obtained by utilizing averages of durations and averages of latencies as features in the SVM. The achieved performance might stem from the mean values reducing the fluctuations in the values, effectively lowering the overall noise that is created by the distortion. This method leads to a lower amount of information being available for analysis, yet the increased robustness in the classification may counterbalance this, advocating for the importance of data quality over quantity.

Sub-Question 3: *Is it feasible to detect whether the timing data collected is distorted?*

The collected timing data has revealed through the measurement of IQR a greater level of inconsistency and a higher range compared to the simulated dataset. These findings highlight the need for further research and development in this domain, particularly in identifying and addressing distorted timing data. Such efforts will be crucial in improving the accuracy and reliability of timing data analysis and interpretation in academic and research settings.

Research Question: *Is it possible to identify soft biometrics, such as age and gender, even when the collected keystroke dynamics data is distorted?*

This study concludes that it is, to a certain degree and context, feasible to detect soft biometrics, particularly age and gender, even when the keystroke timing data is distorted, consistent with the hypothesis. It is shown through experiments that the distortion of data can perform relatively well with the machine learning model Support Vector Machine. Age classification did not reach the same level of performance and accuracy as gender classification, however, the results were not significantly reduced by introducing distorted data.

8.2 Contributions

There have not been extensive studies within the field of distorted KD data in prior literature. The limited availability of directly comparable studies highlights the uniqueness of our research and its potential to make significant contributions to the field. Furthermore, this presents new possibilities for future investigations, which may offer novel perspectives and valuable insights.

This thesis has revealed that the Interquartile Range (IQR) can potentially serve as a method of detecting distortion within data sets. The IQR provides a clear visualization of the middle 50% of values, making it an efficient tool for identifying outliers and measuring the degree of data distortion. This technique proves particularly advantageous in situations where skewed data or extreme values may negatively impact the accuracy of classification. The potential for this approach to be applied to other fields experiencing similar difficulties in identifying distortion underscores its usefulness and versatility.

The research presented has shown that the use of the mean values of latencies and durations can be effective features for classification of soft biometrics, particularly for gender classification with distorted data. By utilizing mean values, it is possible to effectively eliminate noise in the data, leading to improved accuracy for gender-based predictions. Furthermore, this method can be applied to other classification tasks that involve distorted data. Employing this approach can potentially enhance the reliability and validity of the results obtained from such analyses.

Another contribution from this thesis is the exploration of the potential of unsupervised machine learning techniques for classifying soft biometrics. Unsupervised learning approaches offer an alternative to traditional supervised methods, especially when obtaining labeled data is challenging. While initial findings indicate promise, further research is necessary to comprehensively evaluate the strengths and limitations of these methods in the context of soft biometric classification.

8.3 Further work

This section provides several ideas that have been brought to light during the research within KD.

8.3.1 Unsupervised Machine Learning Classifier

A real-time application analyzing KD data would not be able to know if the user is a male or female, child or adult. The approaches for theoretical scenarios and real-time scenarios do not align, as the latter would not have access to a reference table. One solution would be to create a general reference table from a large number of users. However, this requires a vast amount of time and resources.

An interesting approach would be to introduce an unsupervised machine learning classifier. Throughout this study, the potential and increasing importance of unsupervised machine learning methods have been recognized, and it would be beneficial to examine these strategies further.

8.3.2 Explore other Machine Learning models

This study utilizes SVM to classify gender and age, exhibiting potential. However, further investigation is necessary to evaluate and verify the provided results. There are numerous supervised ML models, such as Tree models that have had promising results for classification with standard KD data. Several models would also contribute to understanding how the classification techniques handle distorted data, the generalizability, challenges and further development of efficient classification models.

8.3.3 Continuous Classification on distorted Keystroke Dynamics Data

This study is based on predefined texts, which show potential for the classification of soft biometrics. Future work should examine the performance and effect distorted keystroke timing data has when utilizing free text. Comprehension of the effects of the distortion of the performance of soft biometrics classification in real-world scenarios with variable and unforeseen text will provide valuable insights for future developments. This will further enhance the robustness and applicability of classification techniques of soft biometrics classification.

8.3.4 Continuous Detection of distorted Keystroke Dynamics Data

Continuous detection of distorted KD data could automate the process of handling distorted KD data for authentication and identification. This is specifically applicable to real-world scenarios, creating possibilities for proactive error correction, adaptive modeling, and improved user experiences.

Bibliography

[ACB18] A. Alshehri, F. Coenen, and D. Bollegala, «Iterative keystroke continuous authentication: A time series based approach», *Künstl Intell*, vol. 32, pp. 231–243, 2018.

[Al 19] Al Sweigart, *PyAutoGUI Documentation*, https://pyautogui.readthedocs.io/en/latest/, Accessed: 29.05.2023, 2019.

[AYRA21] N. Alshareef, X. Yuan, K. Roy, and M. Atay, «A study of gender bias in face presentation attack and its mitigation», *Future Internet*, vol. 13, no. 9, 2021. [Online]. Available: https://www.mdpi.com/1999-5903/13/9/234.

[Bon21] K. Bonnerud, «Write like me: Personalized natural language generation using transformers», *Ntnu.no*, 2021. [Online]. Available: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2835483.

[Bou12] P. Bours, «Continuous keystroke dynamics: A different perspective towards biometric evaluation», *Information Security Technical Report*, vol. 17, pp. 36–43, Feb. 2012.

[CGRL20] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, «A comprehensive survey on support vector machine classification: Applications, challenges and trends», *Neurocomputing*, vol. 408, pp. 189–215, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220307153.

[Cha] D. Chakravorty, *Confusion matrix*, https://www.debadityachakravorty.com/ai-ml/cmatrix/, Accessed: 2023-05-23.

[CKH+19] C. Choi, J. Kim, H. Han, D. Han, and H. S. Kim, «Development of water level prediction models using machine learning in wetlands: A case study of upo wetland in south korea», *Water*, vol. 12, no. 1, p. 93, Dec. 2019. [Online]. Available: http://dx.doi.org/10.3390/w12010093.

[CSPK07] K. J. Cios, R. W. Swiniarski, W. Pedrycz, and L. A. Kurgan, «The knowledge discovery process», in *Data Mining: A Knowledge Discovery Approach*. Boston, MA: Springer US, 2007, pp. 9–24. [Online]. Available: https://doi.org/10.1007/978-0-387-36795-8_2.

[Cud] M. Cuda, *Distance metrics*. [Online]. Available: https://numerics.mathdotnet.com/Distance.

[DG04] K. Delac and M. Grgic, «A survey of biometric recognition methods», in *Proceedings. Elmar-2004. 46th International Symposium on Electronics in Marine*, 2004, pp. 184–193.

[Dor18] A. J. Dorca, *Identifying users using Keystroke Dynamics and contextual information*. Feb. 2018. [Online]. Available: https://www.tdx.cat/bitstream/handle/10803/461468/DorcaJosaAleix-Thesis.pdf?sequence=1.

[Epp10] C. Epp, «Identifying emotional states through keystroke dynamics», inse rtURLhereifavailable, M.S. thesis, University of Saskatchewan, Saskatoon, CANADA, 2010.

[FD12] M. Fairhurst and M. Da Costa Abreu, «Using keystroke dynamics for gender identification in social network environment», in *4th International Conference on Imaging for Crime Detection and Prevention 2011 (ICDP 2011)*, IET, 2012.

[For20] W. E. Forum, «Passwordless authentication: The next breakthrough in secure digital transformation», 2020. [Online]. Available: https://www3.weforum.org/docs/WEF_Passwordless_Authentication.pdf.

[Fro21] J. Frost, *Interquartile range (iqr): How to find and use it*, Aug. 2021. [Online]. Available: https://statisticsbyjim.com/basics/interquartile-range/.

[GDPR16] European Parliament and Council of the European Union, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*, May 4, 2016. [Online]. Available: https://data.europa.eu/eli/reg/2016/679/oj (last visited: Apr. 13, 2023).

[GER11] R. Giot, M. El-Abed, and C. Rosenberger, «Keystroke dynamics overview», in *Biometrics*, J. Yang, Ed., Rijeka: IntechOpen, 2011, ch. 8. [Online]. Available: https://doi.org/10.5772/17064.

[JR08] A. K. Jain and A. Ross, «Introduction to biometrics», in *Handbook of Biometrics*, A. K. Jain, P. Flynn, and A. A. Ross, Eds. Boston, MA: Springer US, 2008, pp. 1–22. [Online]. Available: https://doi.org/10.1007/978-0-387-71041-9_1.

[Kec05] V. Kecman, «Support vector machines – an introduction», in May 2005, vol. 177, pp. 605–605.

[Kum20] A. Kumar, *Different types of distance measures in machine learning - data analytics*, Dec. 2020. [Online]. Available: https://vitalflux.com/different-types-of-distance-measures-in-machine-learning/.

[LAS17] Y. Lim, A. Ayesh, and M. Stacey, «Exploring direct learning instruction and external stimuli effects on learner's states and mouse/keystroke behaviours», in *Proceedings - 2016 4th International Conference on User Science and Engineering, i-USEr 2016*, Institute of Electrical and Electronics Engineers Inc., Feb. 2017, pp. 161–166.

[Mag07] I. Maglogiannis, *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies* (Frontiers in artificial intelligence and applications). IOS Press, 2007. [Online]. Available: https://books.google.no/books?id=vLiTXDHr%5C_sYC.

[MFF+15] A. Morales, M. Falanga, J. Fierrez, C. Sansone, and J. Ortega-Garcia, «Keystroke dynamics recognition based on personal data: A comparative experimental evaluation implementing reproducible research», in *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2015, pp. 1–6.

[Moe21a] T. Moe, *Collectkeystrokes: Webpage to collect simulated keystrokes*, 2021. [Online]. Available: https://github.com/tobiasmoe/CollectKeystrokes.

[Moe21b] T. Moe, «I still know who you are! soft biometric keystroke dynamics performance with distorted timing data», May 2021. [Online]. Available: https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2781218/no.ntnu%3Ainspera%3A77286691%3A32312533.pdf?sequence=1.

[Moe21c] T. Moe, *Simulatekeystrokes: A c++ program*, 2021. [Online]. Available: https://github.com/tobiasmoe/SimulateKeystrokes.

[Moo15] P. Moore, *Behavioral profiling: The password you can't change*, Jul. 2015. [Online]. Available: https://paul.reviews/behavioral-profiling-the-password-you-cant-change/.

[MR00] F. Monrose and A. D. Rubin, «Keystroke dynamics as a biometric for authentication», *Future Generation Computer Systems*, vol. 16, no. 4, pp. 351–359, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X9900059X.

[NAMH14] A. N. H. Nahin, J. M. Alam, H. Mahmud, and K. Hasan, «Identifying emotion by keystroke dynamics and text pattern analysis», *Behaviour & Information Technology*, vol. 33, no. 9, pp. 987–996, 2014. [Online]. Available: https://doi.org/10.1080/0144929X.2014.907343.

[Nat17] National Institute of Standards and Technology, «NIST Special Publication 800-12 Revision 1: An Introduction to Computer Security: The NIST Handbook», National Institute of Standards and Technology, Tech. Rep. 800-12 Revision 1, 2017. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-12r1.pdf.

[NFG+22] T. Neupert, M. H. Fischer, E. Greplova, K. Choo, and M. M. Denner, *Introduction to machine learning for the sciences*, 2022.

[Oraar] Oracle Corporation, *Java AWT Robot Class Documentation*, https://docs.oracle.com/javase/8/docs/api/java/awt/Robot.html, Accessed: Date, Year.

[Osc03] P. Oscarson, «Information security fundamentals», C. Irvine and H. Armstrong, Eds., pp. 95–107, 2003.

[Pen17] A. Pentel, «Predicting age and gender by keystroke dynamics and mouse patterns», in *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, ser. UMAP '17, Bratislava, Slovakia: Association for Computing Machinery, 2017, pp. 381–385. [Online]. Available: https://doi.org/10.1145/3099023.3099105.

[Pen19] A. Pentel, «Predicting user age by keystroke dynamics», in Jan. 2019, pp. 336–343.

[Ple22] C. V. Plesner, «Privacy vs. security», Department of Information Security, Communication Technology, NTNU – Norwegian University of Science, and Technology, Project report in TTM4502, Dec. 2022.

[Qin20] T. Qin, «Machine learning basics», in *Dual Learning*. Singapore: Springer Singapore, 2020, pp. 11–23. [Online]. Available: https://doi.org/10.1007/978-981-15-8884-6_2.

[Rat] Ratatype, *What is the average typing speed?*, https://www.ratatype.com/learn/average-typing-speed/, Accessed: 2023-05-28.

[RMT21] J. Rwigema, J. Mfitumukiza, and K. Tae-Yong, «A hybrid approach of neural networks for age and gender classification through decision fusion», *Biomedical Signal Processing and Control*, vol. 66, p. 102 459, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421000562.

[RN09] A. Ross and K. Nandakumar, «Fusion, score-level», in *Encyclopedia of Biometrics*, S. Z. Li and A. Jain, Eds. Boston, MA: Springer US, 2009, pp. 611–616. [Online]. Available: https://doi.org/10.1007/978-0-387-73003-5_158.

[RPK+22] S. Roy, J. Pradhan, A. Kumar, D. Adhikary, U. Roy, D. Sinha, and R. Pal, «A systematic literature review on latest keystroke dynamics based models», *IEEE Access*, vol. 10, pp. 1–1, Jan. 2022.

[RRS18] S. Roy, U. Roy, and D. D. Sinha, «Analysis of typing pattern in identifying soft biometric information and its impact in user recognition», in *Information Technology and Applied Mathematics*. 2018, pp. 69–83.

[SCRB14] S. Z. Syed Idrus, E. Cherrier, C. Rosenberger, and P. Bours, «Soft biometrics for keystroke dynamics: Profiling individuals while typing passwords», *Computers & Security*, vol. 45, pp. 147–155, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404814000893.

[Sha20] P. Sharma, *Understanding distance metrics used in machine learning*, Feb. 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/02/4-types-of-distance-metrics-in-machine-learning/.

[SRF19] S. F. N. Sadikan, A. A. Ramli, and M. F. M. Fudzee, «A survey paper on keystroke dynamics authentication for current applications», *AIP Conference Proceedings*, vol. 2173, no. 1, p. 020 010, Nov. 2019.

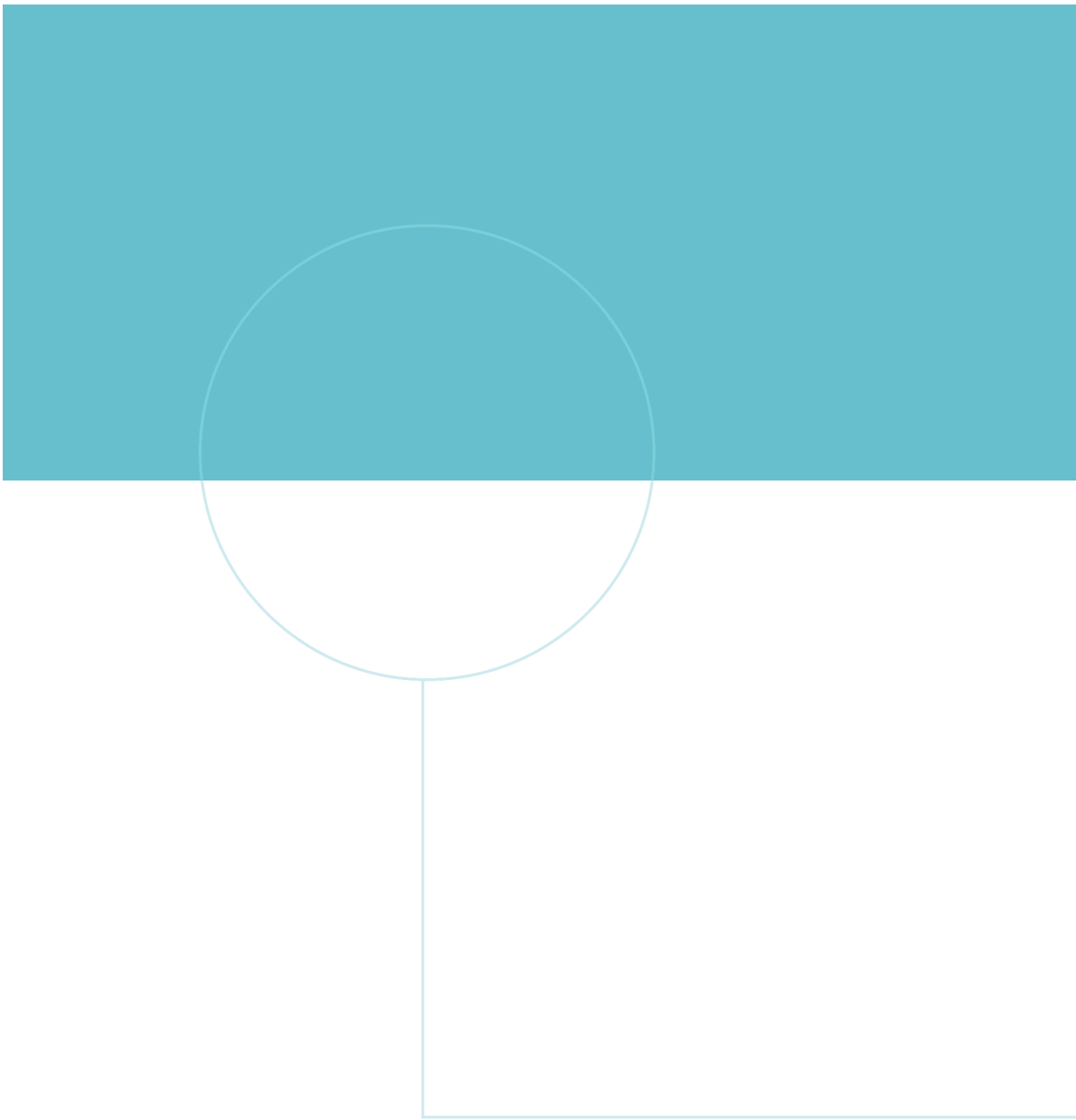
[Sta15] W. Stallings, *Computer Security. Principles and Practice*. UNSW Canberra: Pearson, 2015.

[Sye14] S. Z. Syed Idrus, «Soft biometrics for keystroke dynamics», Dec. 2014.

[TA20] I. Tsimperidis and A. Arampatzis, «The keyboard knows about you: Revealing user characteristics via keystroke dynamics», *International Journal of Technoethics*, vol. 11, pp. 34–51, Jul. 2020.

[Wik22] Wikipedia contributors, *Polynomial kernel — Wikipedia, the free encyclopedia*, [Online; accessed 14-June-2023], 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Polynomial_kernel&oldid=1084678752.

[YLH03] S. Yue, P. Li, and P. Hao, «Svm classification: Its contents and challenges», *Appl. Math. Chin. Univ.*, vol. 18, pp. 332–342, 2003. [Online]. Available: https://doi.org/10.1007/s11766-003-0059-5.



NTNU

Norwegian University of
Science and Technology